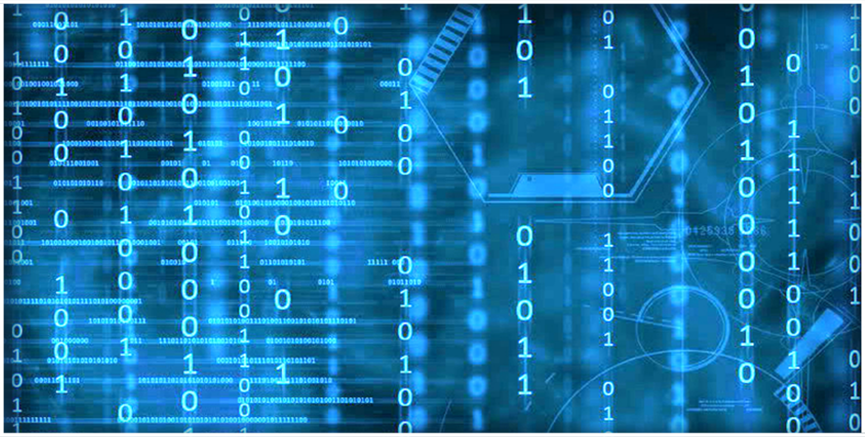
IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 11 Issue 12

December 2020



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

SAI

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

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

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

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

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

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

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

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 11 Issue 12 December 2020
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

CONTENTS

Performance based Comparison between Several Link Prediction Methods on Various Social Networking Datasets (Including Two New Methods)

Ahmad Rawashdeh
Department of Computer Science and Math
University of Central Missouri (UCMO)
Lee’s Summit, MO, USA

***Abstract*—This work extends my previous work on link prediction in Social Networks. In this research, I used two additional datasets, Twitter dataset and Facebook Social Circles Dataset and I ran link prediction methods on these datasets. In my previous work, I performed experiment on the Facebook dataset and proposed two new link prediction methods: Neighbors Connectivity and Common Neighbors of Neighbors (CNN). As in my previous work, in this work, I ran the link prediction methods for several training and testing sizes. Results showed that For Facebook dataset, random had the highest precision, followed by Neighbors Connectivity, then Preferential Attachment, followed by Jaccard/CC, Adamic-Adar, finally CNN. For Twitter dataset, random achieved the highest precision. Preferential Attachment achieved the next highest precision, and Adamic-Adar achieved the least precision. For Facebook Social Circles dataset, Preferential-Attachment achieved the highest precision of 1.08891 followed by random for a training and testing sizes of (1535, 2504) respectively. That is said with slight variation on the orders depending on the training and testing size. The low precision values achieved with Facebook and Twitter datasets are due to the graph types which are sparse as indicated in the datasets websites which confirms Kleinberg finding.**

Keywords—Social networks; link prediction; comparison; experiment

I. Introduction

Social Networks have become an essential part of our lives nowadays. Social Networks’ providers are now competing to offer users the best platform, services, and safe environment for all social activities including social collaboration, networking, sharing of textual, image, or video posts and even reaction to such posts in the form of likes or replies. One of the most popular Social networking sites are Facebook[1] (and the likes including Friendster[2] before changing to entertainment website, Zorpia[3] which is now two, and Myspace[4]), Twitter[5], Instagram[6], YouTube[7], Reddit[8], LinkedIn[9], and some of the new ones such as TikTok[10], Snapchat[11] to name a few. Knowing that every social networking website has a unique as well as common audience with others, some are known to be different than the others in providing certain services. For example, Facebook is used for friendship and content (videos, images) sharing and replies. Twitter is for posting short posts, called tweets, and for using Hashtags which helps in getting the newest posts about a certain important event. Instagram is for images and videos sharing. YouTube is for video sharing (originally was part of what was known as Web 2.0). Reddit is for news. LinkedIn is for professional networking. This work focuses on the two-social networking: Facebook and Twitter. The reason for that follows in the next paragraphs.

Facebook and Twitter remain two of the most popular social networking websites. With 2.70 billion users (see https://www.omnicoreagency.com/facebook-statistics/), Facebook is the largest Social Network to date (see https://makeawebsitehub.com/social-media-sites/). Twitter is the 8th largest (see https://makeawebsitehub.com/social-media-sites/) with users count of 340 million (see https://www.omni coreagency.com/twitter-statistics/). That is enough said about the importance of social networking and their applications. In Facebook, most users are familiar with the “People you May Know” feature which is an example of a link prediction application/service (see People You May Know at https://www.facebook.com/help/www/336320879782850). This work is concerned with this kind of application of social network, namely the application or problem of link prediction in social network. So, what is link prediction?

Link Prediction is still one of the active areas in research due to the importance of its various applications which range from predicting links of friendship/followship in social network such as Facebook and Twitter respectively to areas such as biology, co-authorship, networking, and medicine. Which are only few of the examples of the areas or domains where link prediction could be used. More about the application of link prediction can be found at [1]. The importance of this work can be realized by understanding the importance of applications of link prediction. The reader may

[1] www.facebook.com
[2] www.Friendster.com
[3] www.zorpia.com
[4] www.mysapce.com
[5] www.twitter.com
[6] www.instagram.com
[7] www.youtube.com
[8] www.reddit.com
[9] www.linkedin.com
[10] www.tiktok.com
[11] www.snapchat.com

refer to [2], [3], [4], [5], [6], [7], [8], [9] for detailed applications.

Limitation of this work lies in the relatively small dataset sizes, even though it is still huge, compared to the actual data sizes of social networks. However, this current attempt to use the largest possible dataset sizes, on the current system (16384 MB Memory), was a success.

II. Related Work

So much work has been conducted on social network, particularly on link prediction on social networks. John Kleinberg [10] was the first to term this kind of research as link prediction. He compared between several link prediction methods including, to name a few: Common Neighbors, Jaccard, Adamic-Adar, Preferential Attachment, and random. Also, he considered global link prediction methods such as Kats which uses the ensembles of all paths between the nodes/vertices of interest. Kleinberg ran link prediction algorithms on co-authorship network. Since the prediction algorithms had very low performance values, he measured the relative performance of various predictors versus the common neighbors as well as random predictor (factor improvement). He found that Adamic-Adar and Kats had the highest factor improvement over random.

A survey of link prediction method as well as an experiment was carried out in [11], in which the authors classified link prediction methods into: node based, link based, and path based. The experiment was conducted on the Epinion (a website of reviews) dataset. 10% of the links from the graph were removed for testing purposes. The results showed superiority of Local Random Walk (LRW) algorithm, even though nodes' neighbors, Jaccard, and Supervised Random Walk (SRW) were close. It was concluded that LRW was the best in terms of precision among the compared methods (about 12 methods). However, only one dataset was used, and no new methods were proposed. Also, no detailed information about the used dataset was provided.

In [12] another survey of link prediction methods was performed, but it is theoretical based survey which clearly implies that no experiment was carried out as in this work.

In [13] an experimental comparison of five link prediction methods were performed. The paper also introduced a new link prediction method called LinkGyp. The experiment was carried on three datasets (not the same that are used in this work).

Other works in which more link prediction methods were proposed can be found at [14] and [6] which investigated a machine learning classifier to predict links. The earlier, constructed a feature vector from topological information and node attributes and was evaluated on a co-authorship dataset. The later evaluated the algorithm on 10 different datasets. Also, the research in [15] evaluated two new proposed methods.

In [16] a new link prediction method was proposed, called Time-aware Multi-relational Link Prediction (TMRLP) which combine the dynamic or of the graph topology and interaction history. Results showed that it outperformed the existing methods when ran on DBLP dataset.

An intensive comparison between link prediction methods was conducted in [2]. However, even though the comparison included so many link prediction methods, the datasets used are different from the one used in this work. This work's focus is in using link prediction in Social Network. The research work just been cited found that new links can be better predicted using only local or quasi-local information in most networks. Considering indirect connections only adds noise and computational complexity to the link prediction problem.

All the cited work before differs from this current work in either the used datasets, the application of link prediction, the existing and proposed link prediction methods which are studied, the results, or the experiment setting (running the experiment for several training and testing sizes which has been done in this work).

In [17], I investigated the semantics in link prediction. Also, our work in [18] was about using semantic in finding similarity in Social Networks which can be used in link prediction.

This research extends the work in [1], which compares between link prediction methods and proposes two new methods, by applying the link prediction methods (including the two proposed) on two additional datasets.

III. Problem Definition

In this section, we define formally the problem of link prediction. Given a graph G represented as G = (V, E). Where V is the set of all vertices/nodes, and E is the set of all edges/connection. Which edges might be formed in the future or which missing edges can be predicted? [19] These are two are two different version of the problem. So, given the graph at time t_0, at time t_1, which new edges can be predicted or simply given an instance of the graph, if some edges had been deleted, which missing edges could be predicted. [17].

IV. The Algorithms

The algorithms considered in this paper are: Common Neighbors, Jaccard, Adamic-Adar, Preferential-Attachment, Random, Common-Neighbors-of-Neighbors and Node-Connectivity. The last two methods were proposed by my work at [1].

A. The Formulas for the Algorithms

The formula for links predictions considered in this paper are as follows (more details in [1]):

1. Common neighbors [10]

$$\text{Score (x and y)} = |\urcorner(x) \cap \urcorner(y)| \quad (1)$$

2. Jaccard’s coefficient [10]

$$\text{Score(x and y)} = \frac{|\urcorner(x) \cap \urcorner(y)|}{|\urcorner(x) \cup \urcorner(y)|} \quad (2)$$

3. Adamic-Adar [10]

$$\text{Score (x and y)} = \sum_{z \in |\urcorner(x)| \cap |\urcorner(y)|} \frac{1|}{\log |\urcorner(z)|} \quad (3)$$

4. Preferential attachment [10]

$$\text{Score (x and y)} = |\urcorner(x)| \, . \, |\urcorner(y)| \quad (4)$$

5. Random

Score (user x and user y) =
true or false depending on the binary random value
computed using Math.Random (5)

6. Neighbors Connectivity, Hybrid (proposed in my work in [1])
   a) If Common Neighbors in (1) gives a Score(x and y) >= 1, use that score
   b) Else use the following formula:

Score (x, and y) =
Average degree of neighbors of neighbors of user x +
Average degree of neighbors of neighbors of user y

$$= \frac{\sum_{z \in \urcorner(x)} |\urcorner(z)|}{|\urcorner(x)|} + \frac{\sum_{z \in \urcorner(y)} |\urcorner(z)|}{|\urcorner(y)|} \quad (6)$$

7. Common Neighbors of Neighbors is calculated as follows (proposed in my work in [1])

Score (x and y) =

$$\Big|(\text{neighbors}\big(\text{neighbors}(x)\big) \cap (\text{neighbors}(\text{neighbors}(y))\Big| = \Big|\urcorner\big(\urcorner(x)\big) \cap \urcorner\big(\urcorner(y)\big)\Big| \quad (7)$$

V. Datasets

This work extends my work at [1] by comparing between link prediction on additional datasets. I used Facebook (friendship: undirected) in that work. In this work, I am also using Twitter (fellowship: directed), and Facebook SocialCircles (contains friendship, profiles’ features, and circles), and possibly MoviesGalaxies Social datasets (edges connecting similar movies) in the near future.

I used the Facebook (more about it in my work in [1]), Twitter, and Facebook Social Circles datasets. Information about them can be found in Table I.

More about Twitter dataset can be found and downloaded at the link http://networkrepository.com/ego-twitter.php. The Twitter dataset was created by the work at [20]. The dataset, as indicated on the website, contains followship: user to user following information. A node represents a user. An edge indicates that the user represented by the left node follows the user represented by the right node. It is worth noting that the graph is sparse, with not so many edges (that explains the low precisions listed in the result section which is fine since also Kleinberg got similar low performance values, so he measured the factor improvement over selected algorithm [10]).

TABLE I. INFORMATION ABOUT USED DATASETS (SIZE: NUMBER OF NODES, AND NUMBER OF EDGES)

	nodes	edges	Average clustering coefficient
Facebook	2699 nodes used all used in this and in previous work	2981 edges all used in this and in previous work	0.0272474
Twitter	Total: 23400 nodes used: only nodes whose indices are from 1 to 20000 are included in the processed dataset. For example, the node 20011 was not included.	Total:33101 edges used: < 33101 edges	0.1014
Facebook Social Circles	4039 nodes	edges 88234	0.6055

Facebook Social Circles can be downloaded at the link https://snap.stanford.edu/data/egonets-Facebook.html. The dataset was created from the work in [21]. The dataset contains circles (or friends list from Facebook). Also, the dataset contains node features (profiles), circles, and ego networks.

VI. Running the Algorithms

To know more about the program which I wrote for the experiment, the reader may refer to my previous paper at [1]. I have extended the program and ran it on Facebook dataset (again), Twitter dataset, and Facebook Social Circles Dataset, and then generated output for the three datasets. The precision of every studied link prediction method was calculated for Twitter, Facebook, and Facebook Social Datasets for various training and testing sizes. The link prediction methods used are Common Neighbors, Jaccard, Adamic-Adar, Common Neighbors of Neighbors (CNN), Node Connectivity and Random. CNN and Node Connectivity were proposed by me and experimented with in my previous paper. Several training and testing sizes of the datasets were used to reach the most ever possible generic conclusion based on the current experiment on the datasets. The training sizes attempted are 20%, 25%, 40%, 50%, and 62% of the total dataset size (see Table I). The following were the training and testing sizes for the Facebook dataset in the format of (training, testing): (2159 ,540), (2000,699), (1620 ,1079), (1350 ,1349), and (1000 ,1699) [1]. Initially, for Twitter dataset, I used the following training sizes: 2196, 2745, 4392, 5491, 6808. Then used later the following training sizes and testing sizes (since I reduced total nodes size to 20000) in the format (training size, testing size): (16000, 4000), (15000, 5000), (12000, 8000), (10000, 10000), (7600, 12400). The later sizes of Twitter dataset were used because I encountered an “out of memory” exception (see the sub section below) but still the same percentage of training data sizes were used. The following were the training and testing sizes for the Facebook Social Circles dataset in the format of (training, testing): (3232, 807), (3030, 1009), (2424, 1615), (2020, 2019), and (1535, 2504).

A. Problem: Memory

The twitter data is very large (could not create a two-dimensional array of size more than 20000 x 20000). So, I had to reduce the graph data to include only edge information for source nodes starting from node 1 to node 20000. That explain the exception which I encountered “OutOfMemory” exception during the running of CC algorithm (Common neighbors), so I had to focus on resolving this issue before running all algorithms on the dataset.

Initially, I was allocating memory for more than 20000 entries, however, running into the exception, “OutOfMemory” made me change my coding. So, technically speaking, I had to modify the class Graph.cs (the class file used to store the graph data) to use linked adjacency matrix (GraphLinkedData of type Dictionary<int, List<int>>) to reduce the used memory. I only stored the links without having to allocate wasted memory as in the case of when using the matrix GraphData of type int [,]. That was enough to make it work, and by the time I had a running program that stores the edges’ information in a dictionary (not matrix), the program was ready to run all algorithms. So, lesson learned, only the links need to be stored, and no memory needs to be allocated for any non-existing link, missing link (between pair of nodes). Sine that correction, I had not encountered any similar error. The keys of the GraphLinkedData dictionary were 20000 of count so I succeeded in using most of the dataset.

As mentioned previously, I had to modify the algorithms (CC, jaccard, Adamic-Adar, preferential-attachment, random, CCNeighbors_of_neighbors, NodeConnectivity) to use the linked list adjacency matrix (GraphLinkedData) when using twitter dataset to avoid the “OutOfMemory” exception.

Another related issue is that random took so much time (to initialize the candidate links and made the random prediction). In [1], For Facebook dataset, I was able to generate a random number for every possible edge between a pair of nodes, however for Twitter, since the number of nodes is very large, 20000, I generated the random number for only MaxTestingSize=12400 (which matches the last testing size for running the experiment on Twitter dataset), because that what was needed.

In summary, I found out having enough memory is important sometimes, however, it turns out that it is the limit of the C# programming language which I’m using. However, the program ran with no additional problems after that.

For Facebook Social Circles Dataset, the memory consumption was as follows while running the program (shown is the status of the program and memory):

- Reading the Facebook Social Circles Dataset (memory consumption less than 500MB).
- End reading the Facebook Social Circles Dataset (memory consumption less than 500MB).
- Common neighbors (memory consumption less than 500MB).
- Jaccard coefficient (memory consumption less than 500MB).
- while running preferential (memory 633 MB).
- while running Adamic (memory 860 MB then 915 MB and it started slowing).
- while running random (memory 1000MB speed got back well again).
- while running neighbor’s connectivity (memory 1900 MB then 1700 MB), at this stage the program become very slow in processing the data.

so, I had to stop the program, modify the code then run it once again but only using Common neighbors, Jaccard, Adamic-Adar, Preferential-Attachment, and random. Then start the application again and run it for remaining algorithms (neighbors’ connectivity, common neighbors of neighbors) after reading the dataset. So, the results for Neighbors Connectivity and Common Neighbors of Neighbors were obtained from a different run than from the remaining algorithms. However, it seemed as if the algorithms Neighbors Connectivity and Common Neighbors of Neighbors were slow because of the processing (higher complexity than the other algorithms) and the program ran them slowly for this dataset not because all algorithms were running altogether (required more than 6 hours to finish). Next section discusses the results of the work.

VII. Results and Discussion

After generating the predictions and calculating the precision, one can refer to Table II, Table III, and Table IV which show the precision of link prediction methods on Facebook Dataset, Twitter dataset, and Facebook Social Circles Dataset, respectively. Each table shows the precision of the algorithms for several training and testing sizes. See Section VI for the training and testing sizes used for each dataset.

The precisions are very low. Which confirms the finding by Kleinberg [10], that is due to the nature of the graph, which is sparse, where there are very few edges compared to all possible edges. Possible edges are equal to:

Possible edges

= all possible source nodes × all possible destination nodes

= (total number of nodes × (total number of nodes -1))

I am not counting self-loops, so,

Possible edges (twitter dataset) =

= 20000 × 19999

= 399980000 possible edges.

Table II shows the precision of all studied link prediction methods which were run on the Facebook dataset using several training and testing data sizes. As it can be observed, overall, random has the highest precision, followed by Neighbors Connectivity, then Preferential Attachment, followed by Jaccard/CC, Adamic-Adar, finally CNN. For CNN, the number of decimal digits used to format the display of the precision value was not enough, so even though it shows as 0, it is not actually 0, because there were positive predictions, but very few compared to the overall possible edges to be predicted. The used String Format method wasn’t using enough decimal digits.

As illustrated in Table III, overall, the algorithms ordered from the one with highest precision to the one with the lowest are as follows: Random, Preferential-Attachment, Neighbors-Connectivity, Common-Neighbors-of-Neighbors (CNN), Jaccard/Common Neighbors, and finally Adamic-Adar.

All algorithms were run at once (together in a single execution). The dataset was read only once; the algorithms were run (one by one) and made their predictions. Then, testing data nodes were selected from the original dataset, and finally evaluation was performed, and results were plotted. All automated by on command to the program console. The algorithms were executed on the Twitter dataset for several values for the sizes of training and testing data (see Section VI). Jaccard and Common Neighbors have the same precision values that is may be due to logical error or simply they produced same output.

Just to confirm the ordering of the algorithms in terms of the precision mentioned earlier, and by checking Table III, one can observe the following: For a training data size of 7600, the random was the highest, followed by preferential attachment, followed by Neighbors_Connectivity, then CNN, then Jaccard/CC, finally Adamic-Adar. For a training data size of 10000, the random had the highest precision, followed by preferential attachment, then Neighbors-Connectivity, then CNN, then Jaccard/CC, finally Adamic-Adar. The same is true about other training data sizes.

Random algorithm predicted a link if the number 1 is generated (see [1]), 0 otherwise. Generating a series of binary random number shouldn't exhibit any pattern (that is what completely random mean) unless it is pseudo random. So, there shouldn't be any pattern among all sequence of generated binary values of 0 or 1 (which are used to decide whether a link is predicted or not in this research) and that how it was implemented, see [1]. On a related note, in an ideal situation the chance of predicting a link using random should 50% and the chance of not predicting a link should also be theoretically 50%. And since the graphs in the Twitter dataset [20] and Facebook dataset[12] are sparse graphs as indicated on the datasets websites, random beat other link prediction methods since there is not enough structural information in the graph (few links and the number of neighbors is few and that is what sparse graph means) and these methods are structural. That is not the case with Facebook Social Circles Dataset (as the graph is more dense and less sparse see Table I).

The third dataset, Facebook Social Circles, is not as sparse as the others, so we see higher precision values for all prediction algorithms. The precisions for training and testing data sizes of (1535, 2504), respectively for one algorithm is the highest for all algorithms compared to other training and testing sizes (for the same algorithm). However, among all algorithms, preferential attachment achieved the highest precision of 1.08891, and that occurred for a training and testing sizes of (1535, 2504). The next highest precision is for random for a value of 0.56517 for the same training and testing sizes. The next highest are the remaining algorithms.

Fig. 1, Fig. 2, and Fig. 3 show the precision values for link prediction methods on Facebook, Twitter, and Facebook Social Circles Datasets respectively for the maximum training data size 1000 for Facebook (training: 1000, testing: 1699), training data size of 7600 for Twitter (training: 7600, testing: 12400), and training data size of 1535 for Facebook Social Circles Dataset (training: 1535, testing: 2504). For Facebook Dataset, Neighbors Connectivity produced the highest precision after random, while for Twitter dataset, Random produced the highest precision and all remaining produce low values.

Fig. 4 shows the precision for link prediction methods for the maximum training data size for all datasets. The reader can see that the Facebook dataset produced the least precision values compared with other datasets, followed by Twitter, and finally Facebook Social Circles dataset which produced the highest precision values for Random and Preferential-Attachment link prediction methods.

[12] http://networkrepository.com/ego-facebook.php

TABLE II. PRECISION OF LINK PREDICTION METHODS ON FACEBOOK DATASETS FOR DIFFERENT TRAINING AND TESTING SIZES (TRAINING SIZE, TESTING SIZE)

	Facebook (http://networkrepository.com/ego-facebook.php)				
	←		**increase**	←	
Common Neighbors	(2159, 540)	(2000, 699)	(1620, 1079)	(1350, 1349)	(1000, 1699)
	0.00005	0.00003	0.00001	0.00001	0.00000
Jaccard	(2159, 540)	(2000, 699)	(1620, 1079)	(1350, 1349)	(1000, 1699)
	0.00005	0.00003	0.00001	0.00001	0.00000
Adamic Adar	(2159, 540)	(2000, 699)	(1620, 1079)	(1350, 1349)	(1000, 1699)
	0.00005	0.00002	0.00001	0.00001	0.00000
Preferential Attachment	(2159, 540)	(2000, 699)	(1620, 1079)	(1350, 1349)	(1000, 1699)
	0.00035	0.00021	0.00008	0.00005	0.00002
CNN (Common Neighbors of Neighbors)	(2159, 540)	(2000, 699)	(1620, 1079)	(1350, 1349)	(1000, 1699)
	0.0001	0.0000	0.0000	0.0000	0.0000
Neighbors Connectivity	(2159, 540)	(2000, 699)	(1620, 1079)	(1350, 1349)	(1000, 1699)
	0.01852	0.01128	0.00486	0.00310	0.00194
Random	(2159, 540)	(2000, 699)	(1620, 1079)	(1350, 1349)	(1000, 1699)
	0.02650	0.01615	0.00673	0.00428	0.00276

TABLE III. PRECISION OF LINK PREDICTION METHODS ON TWITTER DATASET FOR DIFFERENT TRAINING AND TESTING SIZES (TRAINING SIZE, TESTING SIZE)

	Twitter (http://networkrepository.com/ego-twitter.php)				
	←		**decrease**	←	
Common Neighbors	(16000, 4000)	(15000, 5000)	(12000, 8000)	(10000, 10000)	(7600, 12400)
	0.00034	0.00008	0.00038	0.00051	0.00096
Jaccard	(16000, 4000)	(15000, 5000)	(12000, 8000)	(10000, 10000)	(7600, 12400)
	0.00034	0.00008	0.00038	0.00051	0.00096
Adamic Adar	(16000, 4000)	(15000, 5000)	(12000, 8000)	(10000, 10000)	(7600, 12400)
	0.00013	0.00003	0.00015	0.00020	0.00038
Preferential Attachment	(16000, 4000)	(15000, 5000)	(12000, 8000)	(10000, 10000)	(7600, 12400)
	0.01132	0.00255	0.01271	0.01727	0.03338
CNN (Common Neighbors of Neighbors)	(16000, 4000)	(15000, 5000)	(12000, 8000)	(10000, 10000)	(7600, 12400)
	0.0006	0.0001	0.0007	0.0009	0.0017
Neighbors Connectivity	(16000, 4000)	(15000, 5000)	(12000, 8000)	(10000, 10000)	(7600, 12400)
	0.00913	0.00204	0.01022	0.01390	0.02688
Random	(16000, 4000)	(15000, 5000)	(12000, 8000)	(10000, 10000)	(7600, 12400)
	0.10981	0.02478	0.12418	0.16908	0.32700

TABLE IV. PRECISION OF LINK PREDICTION METHODS ON FACEBOOK SOCIAL CIRCLE DATASET FOR DIFFERENT TRAINING AND TESTING SIZES (TRAINING SIZE, TESTING SIZE)

	Facebook Social Circles				
	https://snap.stanford.edu/data/egonets-Facebook.html				
Common Neighbors	(3232, 807)	(3030, 1009)	(2424, 1615)	(2020, 2019)	(1535, 2504)
	0.00765	0.01287	0.02808	0.02274	0.06879
Jaccard	(3232, 807)	(3030, 1009)	(2424, 1615)	(2020, 2019)	(1535, 2504)
	0.00765	0.01287	0.02808	0.02274	0.06879
Adamic Adar	(3232, 807)	(3030, 1009)	(2424, 1615)	(2020, 2019)	(1535, 2504)
	0.00765	0.01287	0.02808	0.02274	0.06879
Preferential Attachment	(3232, 807)	(3030, 1009)	(2424, 1615)	(2020, 2019)	(1535, 2504)
	0.12079	0.20386	0.44354	0.35981	1.08891
CNN (Common Neighbors of Neighbors)	(3232, 807)	(3030, 1009)	(2424, 1615)	(2020, 2019)	(1535, 2504)
	not available took so much time				
Node Connectivity	(3232, 807)	(3030, 1009)	(2424, 1615)	(2020, 2019)	(1535, 2504)
	not available took so much time				
Random	(3232, 807)	(3030, 1009)	(2424, 1615)	(2020, 2019)	(1535, 2504)
	0.06255	0.10574	0.23033	0.18669	0.56517

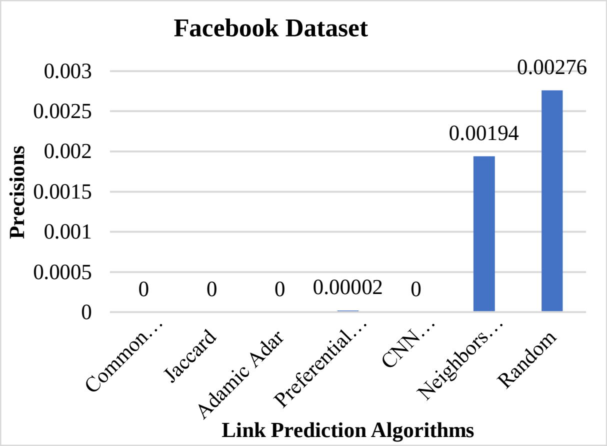


Fig. 1. Precision for the Algorithms on Facebook Dataset for the Max Training Size.

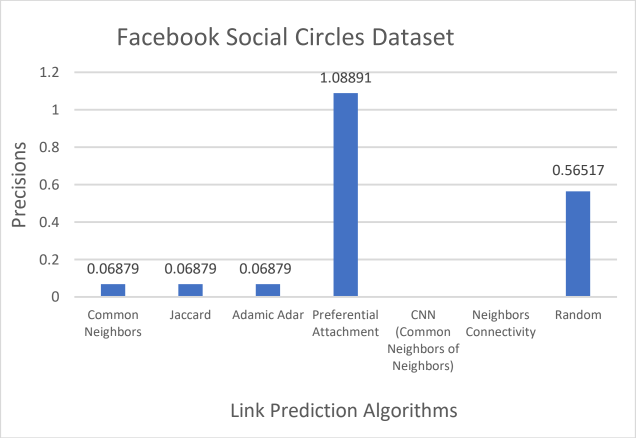


Fig. 3. Precision for the Algorithms on Facebook Social Circles Dataset for the Max Training Size.

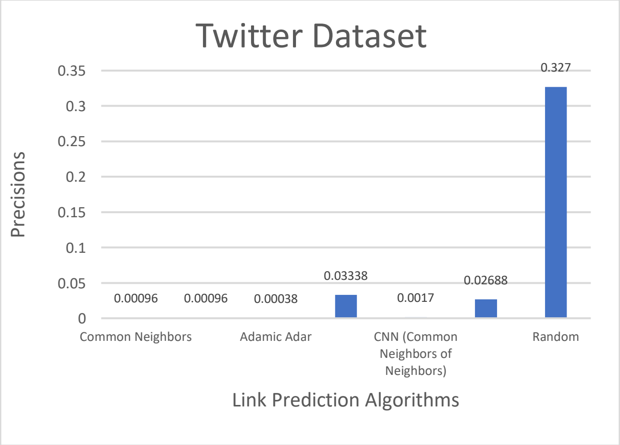


Fig. 2. Precision for the Algorithms on Twitter Dataset for the Max Training Size.

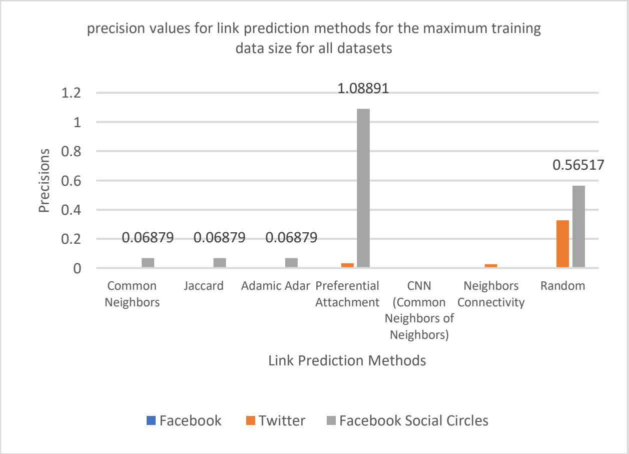


Fig. 4. Precision Values for Link Prediction Methods for the Maximum Training Data Size for All Datasets.

VIII. Conclusion

For Facebook dataset, random had the highest precision, followed by Neighbors Connectivity, then Preferential Attachment, followed by Jaccard/CC, Adamic-Adar, finally CNN. For Twitter dataset, random achieved the highest precision. Preferential Attachment achieved the next highest precision, and Adamic-Adar achieved the least precision. The running time of the algorithms was about (an estimate) 2-4 hours for Twitter and less than an hour for Facebook dataset. For Facebook Social Circles dataset, Preferential-Attachment achieved the highest precision of 1.08891 followed by random for a training and testing sizes of (1535, 2504), respectively.

IX. Future Work

Future work lies in considering new link prediction methods which could achieve better results and focus on factor improvement over certain predictor algorithm, similar to what Kleinberg did who also got low performance values. Also, finding the precision for the remaining methods for Facebook Social Circles Dataset. Another open area for research is using content as well as structural information of the graph in link prediction, and this could be done using the Facebook Social Circles Dataset. Previously, I considered the semantic in link prediction and that is also another very interesting area of link prediction.

References

[1] Rawashdeh A. An Experiment with Link Prediction in Social Network: Two New Link Prediction Methods. In: Arai K, Bhatia , Kapoor , editors. Future Technologies Conference; 10 October 2019. p. 563-581. Available from: https://link.springer.com/book/10.1007/978-3-030-32523-7.

[2] Martínez V, Berzal F, Cubero JC. A survey of link prediction in complex networks. ACM computing surveys (CSUR). 2016;49(4):1-33.

[3] Jalili M, Orouskhani Y, Asgari M, Alipourfard N, Perc M. Link prediction in multiplex online social networks. Royal Society open science. 2017;4(2):160863.

[4] Yang JX, Zhang XD. Revealing how network structure affects accuracy of link predictio. The European Physical Journal B. 2017;90(8):157.

[5] Wang T, He XS, Zhou MY, Fu ZQ. Link prediction in evolving networks based on popularity of nodes. Scientific reports. 2017;7(1):1-10.

[6] Fire F, Tenenboim-Chekina L, Puzis R, Lesser O, Rokach L, Elovici Y. Computationally efficient link prediction in a variety of social networks. ACM Transactions on Intelligent Systems and Technology (TIST). 2014;5(1):1-25.

[7] Al Hasan M, Chaoji V, Salem S, Zaki M. Link prediction using supervised learning. In: SDM06: workshop on link analysis, counter-terrorism and security; 2006. p. 798-805.

[8] Ibrahim NMA, Chen L. Link prediction in dynamic social networks by integrating different types of information. Applied Intelligence. 2015;42(4):738-750.

[9] Dai C, Chen L, Li B. Link prediction based on sampling in complex networks. Applied Intelligence. 2017;47(1):1-12.

[10] Liben-Nowell D, Kleinberg J. The link prediction problem for social network. Journal of the American Society for Information Science and Technology. 2007 March;58(7):1019–1031.

[11] Sharma D, Sharma U, Khatri SK. An Experimental Comparison of the Link Prediction Techniques in Social Networks. International Journal of Modeling and Optimization. 2014;4(1):21-24.

[12] Kushwah AKS, Manjhvar AK. A review on link prediction in social network. International Journal of Grid and Distributed Computing. 2016;9(2):43-50.

[13] Nandi G, Das A. An Efficient Link Prediction Technique in Social Networks based on Node Neighborhoods. International Journal of Advanced Computer Science And Applications. 2018;9(6):257-266.

[14] Liang Y, Huang L, Wang Z. Link prediction in social network based on local information and attributes of nodes. Journal of Physics: Conference Series. 2017;887(012043):10-1088.

[15] Dong L, Li Y, Yin H, Le H, Rui M. The Algorithm of Link Prediction on Social Network. Mathematical Problems in Engineering, vol. 2013. 2013;2013.

[16] Sett N, Basu S, Nandi S, Singh SR. Temporal link prediction in multi-relational network. World Wide Web. 2018;21:395--419.

[17] Rawashdeh A. Semantic Similarity of Node Profiles in Social Networks. Electronic Thesis and Dissertations Center. 2015 119.

[18] Rawashdeh A, Rawashdeh M, D´ıaz I, Ralescu. Measures of semantic similarity of nodes in a social network. In: Laurent , Olivier S, Bernadette BM, Ronald RY, editors. International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems; 2014. p. 76-85. Available from: http://www.lirmm.fr/~lafourcade/pub/IPMU2014/papers/0443/04430076.pdf.

[19] Yang Y, Lichtenwalter RN, Chawla NV. Evaluating link prediction methods. Knowledge and Information Systems. 2015;45(3):751-782.

[20] Rossi RA, Ahmed NK. The Network Data Repository with Interactive Graph Analytics and Visualization. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence; 2015. Available from: http://networkrepository.com.

[21] Leskovec J, Mcauley J. Learning to Discover Social Circles in Ego Networks (Advances in Neural Information Processing Systems 25). 2012.

The Impact of Teaching Operating Systems using Two Different Teaching Modalities

Synchronous Online versus Traditional Face-to-Face Course Delivery

Ingrid A. Buckley
Department of Software Engineering
Florida Gulf Coast University (FGCU)
Fort Myers, Florida, USA

***Abstract*—This paper presents a preliminary look at the performance of two cohorts enrolled in an Operating System course which was taught using two different teaching delivery methods. Operating systems is a technical, senior-level, undergraduate course that includes abstract concepts, mechanisms, and their implementations. This course exposes students to a UNIX-based operating system and includes concurrent programming (threads and synchronization), inter-process communication, CPU scheduling main memory, and virtual memory management. Technical courses present an additional dimension of difficulty when compared to non-technical courses which are more focused on soft skills because they require strong technical skills such as programming and problem-solving. This paper discusses other research studies and statistical data which underscore some of the challenges and differences encountered when teaching a traditional face-to-face versus an online course and the impact on student success. In this work, the 2019 cohort was taught operating systems in the traditional face-to-face modality, while the 2020 cohort was taught the course using the synchronous online modality. The synchronous online modality is very similar to the face-to-face traditional class, in that, lectures are delivered in real-time; this allows students to ask the instructor questions in real-time. Each cohort was tested on the same course objectives (topics) over one semester in 2019 and 2020. The instructor presents the students' performance on three(3) course exams and discusses the differences and similarities in their overall performance between the two groups.**

Keywords*—Operating systems; synchronous online course; traditional course; face-to-face course; online course

I. Introduction

Online education (ELearning) is not a new practice, millions of students have been taking online courses at various universities around the world for over 21 years [15]. It is unlikely for online or ELearning to decrease in the future [3], as the research [20] predicts that the global ELearning market is expected to reach $336.98 billion by 2026. Many studies and research have been completed to understand the difference in student's perception, challenges, and performance in online courses versus a traditional face-to-face course. There are some benefits to offering online courses. They allow a diverse population of students, who are unable to enroll in traditional classes the opportunity to take courses online at their convenience. In recent years, more universities that exclusively offered traditional courses, are now offering online courses to meet the demands and needs of their student body. However, with the advent of the COVID-19 pandemic, many universities were forced to offer traditional courses online for the first time with limited resources and preparation time [6, 17]. Naturally, instructors from a large cross-section of disciplines have different experiences and views about teaching a course online versus face-to-face in a classroom.

Some researchers pose that the swift switch to online, from face-to-face classes with inadequate preparation time, is not technically an online class, instead, they referred to this circumstance as emergency remote teaching (ERT) [12]. Typically, online classes are planned, designed, and constructed specifically for online delivery, and are tailored to suit the nature of the course, whether technical or nontechnical. However, they assert that ERT is different from online learning because it is a temporary shift of instructional delivery to an alternative delivery mode due to crisis circumstances, brought about by war, natural disasters, or a pandemic. In general, some hold a negative view of online classes or online degree programs, assuming that they may not provide the same quality of instruction when compared to traditional face-to-face courses or degree programs. While others believe there is no difference in quality between the two types of instruction delivery methods or student satisfaction [12]. Despite the challenges, discipline, students enrolled in traditional degree programs have a higher chance of completing required courses, whether they are offered online or face-to-face. Students know that completing a required course is a requirement for graduation.

Some studies have been conducted that compare and analyze student retention and performance in online versus traditional face-to-face courses, some of them are discussed in this paper, along with the challenges faced by quickly transitioning from a face-to-face class to an online class. This paper presents a preliminary comparison of student's performance in a traditional operating system course versus the synchronous online version of the course. The students' performance on course exams are examined to determine whether there is a significant difference in teaching a technical computer science course using the synchronous online versus the traditional face-to-face modality.

The outline of the paper is as follows. Section 2 presents a literature review of some studies and statistical data for students enrolled in online and face-to-face courses. Section 3 presents an overview of the traditional and synchronous teaching modalities, learning management systems, and resources used to teach the operating systems course. Section 4 describes the operating systems course structure. Section 5 discusses the approach used to teach the operating system course and highlights the differences and similarities between the face-to-face and the synchronous online class. Section 6 presents the results of the students' performance on three exams and section 7 provide some conclusions.

II. Literature Review

This section provides an overview of some data, and research studies that have been conducted to better understand the various challenges, and factors that may impact student's success in online versus traditional face-to-face courses.

Atchley, Wingenbach, and Aker [5] conducted a study to compare course completion and student performance in online and traditional courses. They found that students performed better in online courses, while students who took classes in traditional courses had higher completion rates. Similarly, Paul and Jefferson [16] completed a similar study, they found that there was no statistical difference in the performance of students in online courses versus face-to-face courses. Cavanaugh and Jacquemin [7] ran a study to determine if there is a difference in grades when students completed courses online versus face-to-face. They found that grades were nearly similar regardless of the method of instruction used. While other studies [19] have found that students taking an online introductory course were more likely to fail compared to students in the face-to-face counterpart of the same course. Christian et. al. [10] did a large-scale study involving 72,000 students enrolled in 433 summer courses. Their result showed that students in the online summer courses performed slightly worse than students in the face-to-face summer courses, and interestingly, the results showed that at-risk students were not at a greater disadvantage in online courses.

Arias, Swinton, and Anderson's [4] study results showed that the face-to-face class performed statistically better than the online class in terms of the exam average and improvement between a pre-test and post-test in a Principles of Macroeconomics course. Aguilera-Hermida [1] surveyed 270 students and found that students preferred face-to-face traditional classes more than online classes. Additionally, Duffin [9] completed an extensive survey involving 1500 colleges to obtain students' perception of the quality of online education in comparison to a face-to-face classroom education in the United States in 2020. This survey showed that 41% of graduate students reported that online college-level education was better than their experiences in college-level face-to-face classroom learning. Gonzalez et. al. [11] completed a study with 458 students and found there was no significant change in student grades on tests, when they learned three subjects in the face-to-face modality, and completed activities in a laboratory, compared to students who learned the same subjects online.

Bozkurt [6] and Salto [17] examined the impact of the COVID-19 pandemic in different countries around the world. More specifically [17] discusses the effect of the pandemic on higher education institutions in Latin America who had to rush to move their teaching online. This study found that these institutions faced two primary limitations: a) the unequal access to technology and the internet; b) their capacity to provide online classes in a short time with limited capacity. Li and Lalani [13] showed that technology and internet access challenges were more common in low-income countries but not only in access but also in the reliability, quality, and speed of the service. Another challenge faced during the COVID-19 pandemic by many academic institutions is the number of students who had access to a working computer [13]. In developed nations, such as Norway, Switzerland, and Austria 95% of students have a computer to use for their schoolwork.

Similarly, in the United States, there is a substantial gap between students from privileged and disadvantaged backgrounds. All most all students from a privileged background reported they had a computer to work on, while nearly 25% of those from disadvantaged backgrounds did not [13]. According to Tam and El-Azar [21], even the software tools used to provide online instruction in developed versus developing nations [8, 14] were also different. They stated that most developed nations utilized a learning management system (LMS) and other technological tools and platforms such as Zoom[1] and Google Drive[2]. While instructors in less developed economies used WhatsApp[3] and e-mail to share lessons and assignments with their students [21].

In traditional face-to-face courses, students are influenced by their peers and they tend to build relationships when they meet regularly each week over one semester or a 4-year degree program. While students taking online courses, may not put effort into interacting with their peers and build a rapport or support system. A good example of these types of online courses is the massive open online course (MOOC)[4]. These are online courses, that allow an unlimited number of students to enroll in a course via the web for free. With MOOCs and other free online courses, students do not have the same financial burden, as they would, if they were enrolled in a traditional or online degree program. Additionally, there is no pressure to complete these types of courses on any schedule, therefore, only highly motivated students generally complete them. Irrespective of the teaching modality, challenges, and limitations, the goals of most universities are to provide quality instruction to their students and achieve high graduation rates.

III. Traditional and Synchronous Online Modalities at Florida Gulf Coast University (FGCU)

Most of the courses are taught face-to-face in the Software Engineering Department at Florida Gulf Coast University (FGCU). The Canvas[5] learning management system (LMS) is used at FGCU to manage courses undergraduate and graduate courses.

[1] https://zoom.us/ (*Zoom footnote*)

[2] https://www.google.com/intl/en_in/drive/ (*Google Drive footnote*)

[3] https://www.whatsapp.com/ (*WhatsApp footnote*)

[4] https://en.wikipedia.org/wiki/Massive_open_online_course (*MOOC footnote*)

[5] https://www.instructure.com/canvas/higher-education (*Canvas LMS footnote*)

A. Canvas Learning Management System (LMS)

Canvas LMS is used to manage course activities (assignments, exercises, quizzes, labs, etc.) and student grades. It allows instructors to assign assignments to students with set due dates. It allows students to submit assignments, project work, exercises and it can be used to administer quizzes and exams. The Canvas LMS is integrated with Canvas Conference which can be used to deliver virtual lectures and virtual office hours. It allows the instructor to upload PowerPoint presentations, work out problems on electronic whiteboards, communicate with students, and answer their questions in real-time. Canvas Conference allows the instructor to record all lectures, which provides students with an opportunity to review lectures as many times as they want.

Additionally, the Canvas LMS is integrated with the LockDown Browser-Respondus[6] and Respondus Monitoring[7] to facilitate exams and quizzes. The LockDown Browser locks down the testing environment within a learning management system. While Respondus Monitoring extends the capabilities of the LockDown Browser, by using a student's webcam and video analytics to prevent cheating during non-proctored online exams.

B. Florida CyberHub

The Florida CyberHub[8] is a virtual learning laboratory that facilitates and enhances cybersecurity educational programs throughout the state of Florida. It provides cloud-based tools, for education and research to faculty members at universities in the State University System of Florida Institutions. The Florida CyberHub installed and set-up virtual machines with a Unix-based operating system and tools to allow students in the operating systems course to learn and investigate operating systems concepts in a remote Sandbox.

C. Traditional Modality

Courses offered in the traditional modality are taught face-to-face in a computer lab where each student has access to a computer. The lab has two projectors on either side of the room. The walls of the lab have many whiteboards for students to practice questions, work individually, or in teams. The front of the lab, has a podium, a whiteboard, a computer, and other teaching tools for the instructor to utilize while teaching the class. Students are required to attend traditional classes in person and these lectures are not recorded. Students use the Canvas LMS to submit assigned classwork and take quizzes during class time. Office hours are given in person at set hours each week.

D. Synchronous Online Modality

Courses taught in the synchronous online modality allows students to attend lectures online, in real-time, at set times each week. With this modality, lectures are taught using Canvas Conference which is integrated with the Canvas LMS. However, students are expected to meet the following minimum equipment and technology requirements to enroll in online (synchronous or asynchronous) courses:

- Processor: Current generation Intel Core Series (i3, i5, i7, i9) or AMD Ryzen equivalent.
- Memory: 8GB RAM.
- Storage: 250GB hard drive/SSD.
- Webcam, microphone, and speakers.
- Wireless internet (WiFi).
- Windows 10 or Mac OS X (High Sierra or newer recommended).
- High speed Internet access at home (10mbps per device is a good rule of thumb).

Operating systems is a technical course that is taught at the undergraduate level, in the traditional face-to-face modality within the Software Engineering Department at Florida Gulf Coast University. However, in 2020, due to the COVID-19 pandemic, this course was offered for the first time, in fall 2020 in the synchronous online modality.

At FGCU, the operating systems course has been taught by the same instructor since 2015 in the traditional face-to-face modality. Operating System is a required course that students in the Software Engineering degree program must complete to graduate and obtain a degree in Software Engineering. This course is offered once, each academic year, in the fall. Seniors typically take the operating system course and are generally more committed and motivated to complete and pass the course irrespective of the modality of the course.

In fall 2019, the operating system course size was capped at 32 students for the traditional face-to-face modality. While in fall 2020, the class size was capped at 42 students for the synchronous online modality. The next section describes the structure of the operating system.

IV. The Operating Systems Course Structure

The operating systems course introduces components of operating systems including process management, memory management, CPU scheduling, threads, synchronization, and protection/security are explored. Contemporary design issues and current directions in the development of operating systems are discussed and case studies of several prominent operating system implementations are studied. The course objectives are:

1) Be able to use some basic Unix commands.

2) Understand the concepts of process and threads and be able to create them in C/C++.

3) Understand and be able to compare different CPU scheduling methods.

4) Understand and explain semaphores and monitors to avoid race conditions in programming.

5) Be able to explain deadlocks and their prevention.

6) Be able to explain the features of segmentation and paging memory management and storage management.

[6] https://web.respondus.com/he/lockdownbrowser/ (*LockDown Browser-Respondus footnote*)

[7] https://web.respondus.com/he/monitor/ (*Respondus Monitoring footnote*)

[8] https://floridacyberhub.org/ (Florida CyberHub footnote)

The prerequisites for the operating systems course are Data Structures and Computer Organization and Assembly Language Programming. Seniors take the operating systems course in the final year of the 4-year Software Engineering degree program.

Seniors are expected to have a good understanding of programming principles and data structures, in addition to an understanding of the organization and low-level computing hardware components, because these principles are fundamental to understanding operating systems principles.

Students are given a variety of activities, exercises, labs, quizzes, and exams in one semester. Exams and quizzes account for the majority of the overall course grade. In 2019 and 2020 the exams and quizzes accounted for 85% and 81% of the overall course grade respectively. The grading scale and is used for both years is shown in Table I.

TABLE I. THE COURSE GRADING SCALE

GRADE	PERCENTAGE
A	93 – 100
A-	91 – 92.9
B+	87 – 90.9
B	82 – 86.9
B–	80 – 81.9
C+	77 – 79.9
C	72 – 76.9
C–	69 – 71.9
D	60 – 68.9
F	0 – 59.9

The textbook used in this course is Operating Systems Concepts, the 9th Edition by Silberschatz, Galvin, and Gagne [18]. This textbook covers each topic in-depth and provides many conceptual and technical practice questions for students.

V. THE APPROACH USED TO TEACH OPERATING SYSTEMS COURSE IN 2019 AND 2020

Generally, seniors in their final year of the Software Engineering degree enroll to the take Operating Systems course each academic year. In fall 2019, the operating systems course was taught face-to-face in the traditional face-to-face modality. The lectures are 75 minutes and are given twice each week. Traditional class lectures are not recorded. All exams and quizzes are given during class time and are invigilated by the instructor. Each academic year, there are two sections of the course. In 2019, both course sections combined had 62 students in total. However, throughout the semester one student dropped out of the course, leaving 61 students. Office hours are given in person, allowing students to drop in at will to get assistance from the instructor.

While, in fall 2020, the operating systems course was taught in the synchronous online modality. Students met with the instructor online twice each week, in real-time, for 75 minutes' lectures using the Canvas Conference tool. All exams are given online, in real-time in Canvas, using Respondus LockDown Browser and Respondus Monitor. Two sections of the course were available, which had had 81 students in total. In the fall 2020 semester, four(4) students dropped the course, leaving 77 students. Office hours were given online using Zoom [24]. However, students had to schedule a time to meet with the instructor for office hours to obtain tutoring instead of dropping in at will, which is more common with the face-to-face office hours.

A. *Operating Systems Exam Format*

The exam formats used in 2019 and 2020 are different. In the traditional class, students took two paper-based exams worth 70% of their overall grade. Additionally, students completed six quizzes worth another 15% of their overall course grade. More specifically, the six quizzes served as the third exam, which means that all exams and quizzes combined are worth 85% of the overall course grade in 2019. The two exams were hand-written. While the six quizzes were completed in the classroom using the Canvas learning management system (LMS) software. In 2019, all exams and quizzes were completed inside the classroom and invigilated by the instructor.

Students in the synchronous online class completed three exams which are worth 81% of their overall course grade. Students completed the three exams synchronously online, using the Canvas LMS, Lockdown Browser Respondus, and Respondus Monitor. These exams had a combination of multiple-choice questions, true/false, fill in the blanks, and questions where students had to show the details of working out a problem. The instructor moderated all online exams, using the moderating feature in the Canvas LMS, to ensure that any issues being experienced by students can be resolved promptly.

In 2019, students would receive their graded exams in hardcopy format, while in 2020 all graded exam comments were accessible online in the Canvas LMS. Irrespective of the teaching modality, all students in each course section are required to take exams and quizzes at the same time to reduce the likelihood of cheating.

B. *Assignments, Activities, Exercises and the Unix Operating System*

In 2019 and 2020, students accessed and submitted assigned work (assignment, exercises and, activities) via the same means. The instructor provided all feedback on assigned work using the Canvas LMS. Most of the students enrolled in the operating system course have Windows operating systems installed on their personal computers. As a result, the instructor collaborated with the Florida CyberHub to provide virtual machines with the Ubuntu operating system, and all the necessary tools required for students to complete their assigned work.

Students are given a unique account and access the Florida CyberHub system remotely from anywhere via the Internet. This allowed students to access a virtual machine with the Ubuntu operating system and to complete assigned labs related to threads, processes, and process synchronization. This worked out well in both the traditional face-to-face and synchronous online classes. In 2019, even though the class was offered in the traditional modality, the CyberHub was used, because it provided convenience and the flexibility for students to continue their work from home without having to go to a lab at the university or to install a virtual machine. In previous years, students installed the virtual machine on their personal computers. However, some students have older computers with a variety of specifications, and some students experienced a lot of technical issues with the installation process. To avoid this problem one system was made available for all students. Additionally, the Florida CyberHub provides technical support to students and instructors, Mondays – Fridays from 8:00 AM to 5:00 PM.

VI. The Traditional (2019) and Synchrous Online (2020) Exam Results

Students enrolled in the operating system course completed three exams over one semester. The results of all exams were taken from the Canvas LMS which stores the grades for all students in fall 2019 and 2020. The students in 2019, took two exams, and six quizzes. Note that, the six quizzes were combined to form (create) the third exam. While students in fall 2020 took three exams. The results of the exams are given in Table II, which provides the averages of each exam and the corresponding standard deviations.

Fig. 1, 2, and 3 provide a comparison of the grade distribution for exams#1, exam#2, and exam#3 in 2019 and 2020 using the grading scale presented in section 4.

The results show that students in the traditional class performed better on exam#1 which covered course objectives 1-3. The students in both cohorts performed relatively the same on exam#2. Exam#2 covered course objectives 3-6. Students in the traditional class did not perform as well on exam#3 as the 2020 cohort. In the traditional class, exam#3 covered all the course objectives (1-6).

The 2020 cohort, performed poorly on exam#1 with an average of 63%, while the average on exam#1 was 82% in the traditional face-to-face class. The results show that the performance on subsequent exams and quizzes did not increase in the traditional class. It is noteworthy to mention that, all the exams in the synchronous online course covered fewer course objectives than the exams in the face-to-face class. This means that the exams in the traditional face-to-face class were more rigorous when compared to the synchronous online class.

TABLE II. RESULTS OF THE THREE EXAMS

	Exam#1	Exam#2	Exam#3
Year: 2019 Traditional Modality			
Number of Students	61	61	61
Average	82.04	75	72.01
Standard Deviation (SD)	9.66	10	7.50
Year: 2020 Synchronous Online Modality			
Number of Students	77	77	77
Average	63.12	72.07	88.60
Standard Deviation (SD)	14.64	15.55	9.99

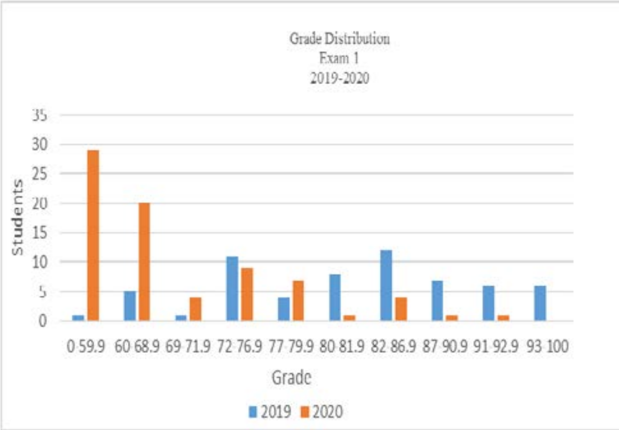


Fig. 1. Exam#1 Grade Distributing.

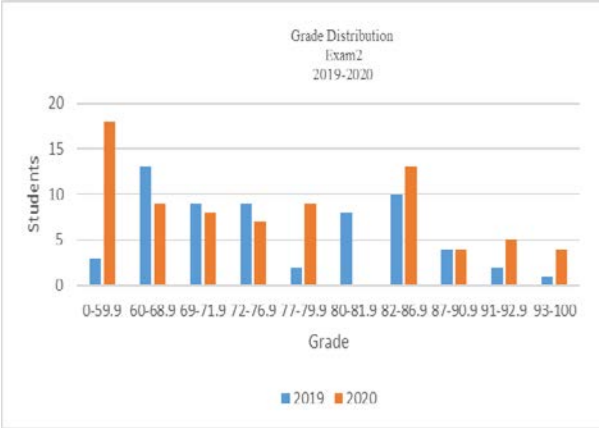


Fig. 2. Exam#2 Grade Distribution.

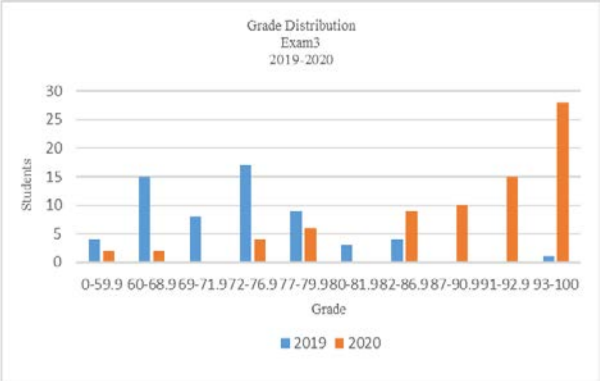


Fig. 3. Exam#3 Grade Distribution.

The student’s in the synchronous online class grades improved steadily after exam#1, obtaining average grades of 72% and 88% on exam#2 and exam#3 respectively. Exam#2 covered course objectives (2 and 3), and exam#3 covered course objectives (5 and 6). Since these exams covered fewer course objectives on each exam, this naturally reduces the complexity of these exams when compared to the exams taken by students in the traditional face-to-face class.

A. *Primilinary Two-tailed Exam Results*

A two-tailed t-test was performed to determine if there was a significant difference between the students' performance on exam#1, exam#2, and exam#3 in the traditional face-to-face versus the synchronous online class. In this test, a p-value less than 0.05 is statistically significant ($p < 0.5$). The preliminary results are given in Table III. The two-tailed t-test p-values show a statistical significance in the students' performance on exam#1 and exam#3. While the p-value for exam#2, shows no statistical significance in the student's performance on exam#2.

The results for the synchronous online modality show an improvement in student's grades from exam#1 to exam#3 over one semester. Students became more motivated after performing poorly on exam#1 because they did not want to risk failing the course. Overall the synchronous online modality yielded better student performance on exams when compared to the traditional face-to-face modality. However, additional studies must be completed using both modalities, to better determine if the synchronous online modality will yield improved student performance when compared to the face-to-face modality when teaching operating systems. Further studies will help determine what other factors could have caused one group to do better on exam#2 and exam#3.

Perhaps student preparedness [2, 22], the number of course objectives covered on each exam, the complexity of the exam questions, when exams are given have a greater influence on students' performance on exams, than the course modality. In general, students do not perform well on quizzes when compared to exams. Students in 2019 completed six quizzes without the assistance of cheat-sheets, and this formed exam#3 for that cohort. While students in 2020 utilized cheat-sheets during exam#3. Furthermore, the weight of exams in 2020 was less, each accounting for 27% of the overall course grade. While in 2019, exam#1 and exam#2 each accounted for 35%, and exam#3 (comprised of 6 quizzes) accounted for 15% of the overall grade.

Additionally, student retention is another primary goal of most universities. However, this presents more of a challenge for students taking asynchronous online courses. Typically, when students get to the senior year in the Software Engineering degree program, they generally graduate and are less likely to abandon or stop their degree program at this late stage. Even if a senior fails a required course, they can take the course at another institution in the following spring or summer, which is in the same academic year, and transfer those credits to FGCU. As a result, failing a required course does not necessarily delay graduation significantly. Furthermore, the majority of the students who took operating systems in fall 2019 completed their required courses and graduated. While students in the 2020 cohort are on schedule to graduate in 2021. Note that the COVID-19 pandemic may impact the 2021 graduation for a variety of personal reasons. However, it is unlikely that the teaching modality will be the primary cause.

TABLE III. PRELIMINARY RESULTS OF THE TWO-TAILED T-TEST

p < 0.5	**Exam#1**	**Exam#2**	**Exam#3**
T-test (p-value)	9.70128E-15	0.204479332	7.96E-20

VII. CONCLUSION

Students who are enrolled in a degree program are more motivated to complete their course because it is a requirement for graduation. It is common, for most universities to offer a mix of traditional and online courses. However, courses taught using the synchronous online approach is quite similar to a traditional face-to-face course. In the former, students meet the instructor online, in real-time at a set time, while students in traditional courses meet their instructors in person in a classroom at a set time for lectures. The main difference between these two modalities is how examinations are administered to students in the synchronous online approach when compared to the traditional course approach. Students take their exams in the classroom with the instructor present, and the instructor serves as the invigilator, which has its limitations. While students in the synchronous online course, take their exam online using monitoring tools that proctor them during the entire examination. Based on the results of the students' performance on three exams, and the other factors that could have contributed to their performance on these exams, the t-test results show a notable steady improvement in students' performance on each successive exam in the synchronous online course, while the face-to-face does not. Additional studies must be undertaken to include additional cohorts, to gather more data which will provide better insights on the impact of teaching operating systems in the synchronous online versus the traditional approach.

REFERENCES

[1] A. P. Aguilera-Hermida, "College students' use and acceptance of emergency online learning due to COVID-19", International Journal of Educational Research Open, vol. 1, 2020, pp.100011, https://doi.org/10.1016/j.ijedro.2020.100011.

[2] A. Alghamdi, A. C. Karpinski, A. Lepp, J. Barkley, "Online and face-to-face classroom multitasking and academic performance: Moderated mediation with self-efficacy for self-regulated learning and gender", Computers in Human Behavior, vol. 102, 2020, pp.214-222, https://doi.org/10.1016/j.chb.2019.08.018.

[3] W. Ali, "Online and remote learning in higher education institutes: A necessity in light of COVID-19 pandemic", Higher Education Studies, vol.10, 2020, no.3, E-ISSN 1925-475X.

[4] J. Arias, J. Swinton, K. Anderson, "Online Vs. Face-to-Face: A Comparison of Student Outcomes with Random Assignment", vol. 12 (2), 2018, pp 1-23, https://files.eric.ed.gov/fulltext/EJ1193426.pdf.

[5] W. Atchley, G. Wingenbach, and C. Aker, "Comparison of course completion and student performances through online and traditional courses international", Review of Research in Open and Distance Learning, vol. 14(4), 2013, pp. 104-116, https://doi.org/10.19173/irrodl.v14i4.1461.

[6] A. Bozkurt, I, Jung, J. Xiao, V.Vladimirschi, et al., "A global outlook to the interruption of education due to COVID-19 pandemic: Navigating in a time of uncertainty and crisis", Asian Journal of Distance Education 15, vol. 1, 2020, pp.1–126.

[7] J. Cavanaugh, and S. Jacquemin, "A Large Sample Comparison of Grade Based Student Learning Outcomes in Online vs. Face-to-Face Courses", Online Learning, 2015, DOI:10.24059/OLJ.V19I2.454.

[8] F. Christian Fischer, X. Di, R. Fernando, D. Kameryn, W. Mark, "Effects of course modality in summer session: Enrollment patterns and student performance in face-to-face and online classes", The Internet and Higher Education, vol. 45, 2020, pp. 100710, ISSN 1096-7516.

[9] R. Dlamini, F. Nkambule, "Information and communication technologies' pedagogical affordances in education", Encyclopedia of Education and Information Technologies. Springer, vol. 1, 2019, pp. 1–14, https://doi.org/10.1007/978-3-319-60013-0_216-1.

[10] E. Duffin, “Opinions of online college students on quality of online education U.S. 2019”, Statista Article, October 26, 2020, https://www.statista.com/statistics/956123/opinions-online-college-students-quality-online-education/.

[11] T. Gonzalez, M. de la Rubia, K. Hincz, M.C. Lopez, L. Subirats, S. Fort, et al., “Influence of COVID-19 confinement in students’ performance in higher education”, EdArXiv, April 20, 2020, https://doi.org/10.35542/osf.io/9zuac.

[12] C. Hodges, S. Moore, B. Lockee, T. Trust, and A. Bond, "The Difference Between Emergency Remote Teaching and Online Learning", EduCauseReview Article, March 27, 2020, https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learning.

[13] C. Li, and F. Lalani, "The COVID-19 pandemic has changed education forever. This is how", World Economic Forum Article, April 29, 2020. https://www.weforum.org/agenda/2020/04/coronavirus-education-global-covid19-online-digital-learning.

[14] C. B. Mpungose, "Emergent transition from face-to-face to online learning in a South African University in the context of the Coronavirus pandemic", Humanities and Social Sciences Communications, vol. 7, 2020, pp. 113. https://doi.org/10.1057/s41599-020-00603-x.

[15] C. Pappas, "Top 20 eLearning Statistics For 2019 You Need To Know [Infographic], Elearning Industry Article, September 24, 2019, https://elearningindustry.com/top-elearning-statistics-2019.

[16] J. Paul and F. Jefferson, "A Comparative Analysis of Student Performance in an Online vs. Face-to-Face Environmental Science Course From 2009 to 2016", Frontiers in Computer Science, vol. 1(7) Digital Education, 2019, pp.1-7, DOI: 10.3389/fcomp.2019.00007.

[17] D. Salto, "COVID-19 and Higher Education in Latin America: Challenges and possibilities in the transition to online education", Elearn Magazine, September 2020, https://elearnmag.acm.org/featured.cfm?aid=3421751.

[18] A. Silberschatz, P. B. Galvin, G. Gagne, “Operating Systems Concepts ”, 9th edition, Hoboken, NJ: Wiley, ISBN-10: 1118063333.

[19] S. B. Waschull, “The online deliver of psychology courses: Attrition, performance, and evaluation”, Teaching of Psychology, vol.28, 2001, pp.143–147.

[20] Syngene Research, “Global E-Learning Market Analysis 2019”. Report, March 2019, https://www.researchandmarkets.com/reports/4769385/global-e-learning-m.

[21] G. Tam and D. El-Azar, "3 ways the coronavirus pandemic could reshape education",https://www.weforum.org/agenda/2020/03/3-ways-coronavirus-is-reshaping-education-and-what-changes-might-be-here-to-stay, April 2020, [Accessed December. 19, 2020].

[22] Y. Tzu-Chi, "Impacts of Observational Learning and Self-regulated Learning Mechanisms on Online Learning Performance: A Case Study on High School Mathematics Course," 2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT), Tartu, Estonia, 2020, pp. 194-197, DOI: 10.1109/ICALT49669.2020.00063.

Reducing Energy Consumption in Microcontroller-based Systems with Multipipeline Architecture

Cristian Andy Tanase
Faculty of Electrical Engineering and Computer Science
Stefan cel Mare University of Suceava
Str. Universitatii 13, 720229 Suceava, Romania

***Abstract*—Current mobile battery powered systems require low power consumption as possible without affecting the overall performance of the system. The purpose of this article is to present a multi-pipeline architecture implemented on a RISC V processor with 4 levels pipeline. Each thread has an assigned CLKSCALE registry that allows to use a clock with a lower or higher frequency, depending on the value written in the CLKSCALE registry. Depending on the importance and the need to be executed at a lower or higher speed each thread will enter into execution with its frequency given by CLKSCALE. It is known that each system has its own "real time". The notion of real time is very relative depending on the environment in which the system operates. Thus, if the system responds to external stimulus for a time that does not affect the operation of the whole, then we say the system is in real time. The system response can be quick or slow. It is important that this response does not lead to malfunction in operation. Therefore, certain threads can work at lower frequencies (those responding to slower external stimulus) and others must operate at high frequencies to allow quick response to fast external stimulus. It is known that the consumed power is directly proportional to the frequency of computing. Thus, the threads that do not require to run at maximum frequency, will consume less energy when they run. The entire system will consume less energy without affecting its performance. This architecture was implemented on a Xilinx FPGA ARTY A7 kit using the Vivado 2018.3 development tools.**

***Keywords*—Multi-pipeline register; RISC V (Reduced Instruction Set Computer); power consumption; multi-threading; FPGA (Field Programmable Gate Array); variable frequency**

I. Introduction

With the development of integrated systems, energy consumption has become a more important constraint in the RTL design. As these integrated systems become more sophisticated, they also need a higher level of performance. The task of satisfying the energy consumption and performance requirements of these embedded systems is a rather difficult task to ensure. One of the most used techniques of enhancing CPU performance is the use of ILP (Instruction Level parallelism) through pipeline technology. The instruction pipeline allows an increased clock frequency by reducing the amount of work to be performed for an instruction in each clock cycle [1].

A reduced energy consumption can increase the standby-time of the terminal which reduces the user annoyance related to recharging the battery too often. A reduced energy consumption could also mean that one can get the same standby-time as earlier but with a smaller sized battery, which reduces the overall size and weight of the terminal. A smaller battery is beneficial from an environmental aspect as well.

Energy consumed in CMOS devices is a product of time and power consumption and is measured in Joules (eq. 1). Power consumption in a CMOS device is consumed both statically and dynamically. Currently, the majority of the power is consumed dynamically, but devices implemented using future process technologies will most likely consume as much static as dynamic power consumption due to increased leakage currents (eq. 2). As can be seen from eq. 3, dynamic power is a function of the voltage level (V_{dd}), frequency (f), capacitive load (C), and the activity factor (α). The activity factor represents the number of transitions between a logic zero and a logic one, which corresponds to charging of capacitances. One observation is that a near-cubic reduction of dynamic power consumption can be achieved by reducing the voltage and frequency. Dynamic power consumption can also be lowered by reducing or eliminating the transistor switching activity. Another source of dynamic power consumption is the current dissipated from short circuits in transistors during switching. Short circuit currents are dissipated when a logical value of a transistor is in the process of doing a transition of its output. During this transition there is a small period of time, where there is a direct path from the supply voltage and the ground, which results in dissipated currents (see eq. 4) The static power consumption comes from non-ideal switch behaviour of transistors, thus the transistors leak currents (see eq. 5) [2] -[12].

$$Energy = Power \cdot Time \tag{1}$$

$$P_{total} = P_{switch} + P_{shortcircuit} + P_{leakage} \tag{2}$$

$$P_{switch} = V_{dd}^2 \cdot f \cdot C_L \cdot \alpha \tag{3}$$

$$P_{shortcircuit(sc)} = V_{dd} \cdot I_{sc} \tag{4}$$

$$P_{leakage} = V_{dd} \cdot I_{leakage} \tag{5}$$

II. Related Work

In [13] the authors propose an runtime environment for next-generation dual-core MCU platforms. These platforms complement a single-core with a low area overhead, reduced design margin shadow-processor. The runtime decreases the

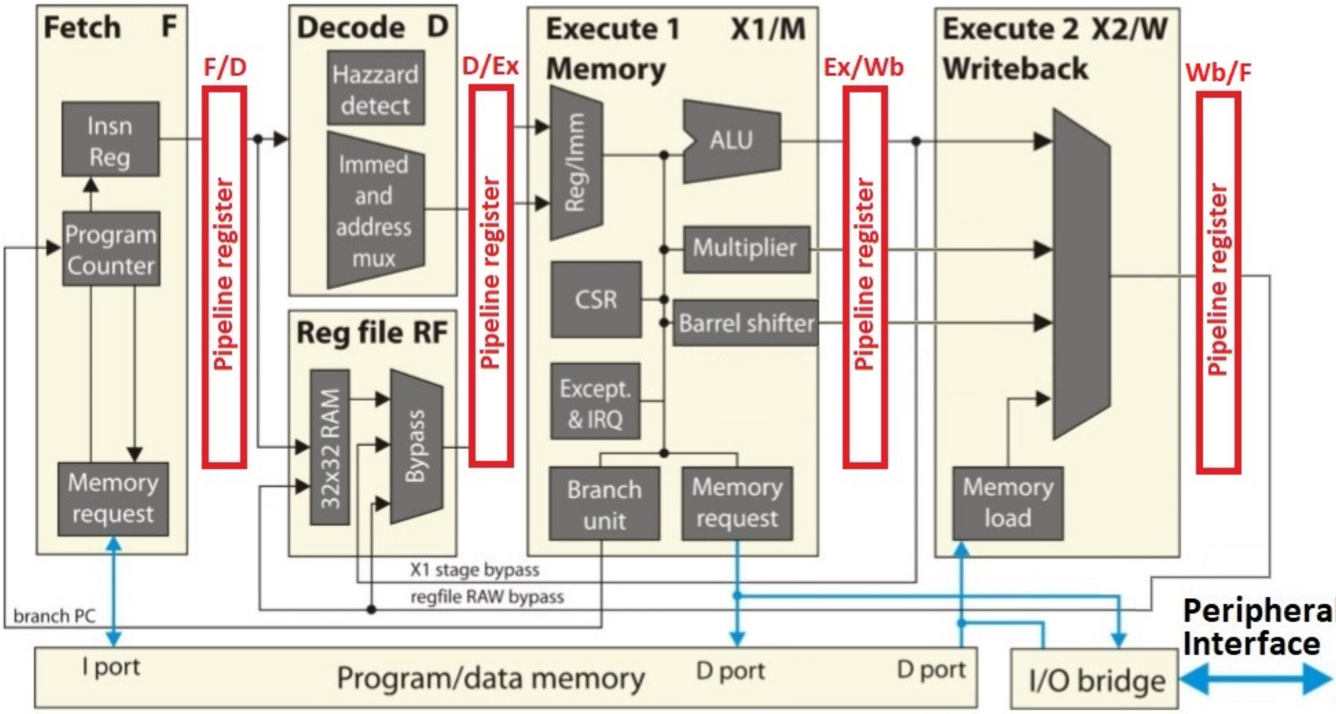


Fig. 1. RISCV Processor (Block Diagram) [17]

overall energy consumption by exploiting design corner heterogeneity between the two cores, rather than increasing the throughput. This allows the platform's power envelope to be dynamically adjusted to application-specific requirements. Depending on the ratio of core to platform energy, total energy savings can be up to 20%.

In [14] the authors realized a single-ISA (Instruction Set Architecture) heterogeneous multi-core architecture, including four Alpha cores and a MIPS R4700 (Microprocessor without Interlocked Pipelined Stages). The allocation of tasks among cores is integrated as part of the operating system.

In [15] the authors show that at NTC (Near-Threshold Voltage Computing - the supply voltage is only slightly higher than the transistor's threshold voltage), is a promising approach to reduce the energy per operation and a simple chip with a single V_{dd} domain can deliver a higher performance per watt than one with multiple V_{dd} domains.

Compared to high-end systems, there has been very little attention paid to task allocation/scheduling on low-cost, limited performance systems. The closest work in this domain would be [16], which focuses on optimal resource management for control tasks in MCUs using a minimal real-time kernel.

III. Background

The proposed architecture was implemented on a RISC V core with four levels pipeline, presented in Fig. 1 [17].

RISC V employs a modified Harvard architecture: code and data reside in a shared 32-bit memory space, but are accessed through separate memory interfaces. Instructions are executed by a four-stage, single-issue pipeline, shown in Fig. 1 and consisting of the following stages:

1) Fetch(F), calculating the address of the next instruction and requesting it from the code memory;
2) Decode(D), which also computes the immediate values (sign-extensions and bit reordering), pre-computes the operand values for the Execute1/Memory(X1/M) stage and manages the hazards by inserting empty instructions into the Execute stage;
3) Execute1/Memory(X1/M), which executes most of the instructions, generates memory addresses and issues memory read/write requests;
4) Execute2/Writeback(X2/W), which completes execution of Load, Multiply and Shift instructions and writes the result back to the CPU registers.

In parallel to the D stage, there is a Register File (RF), hosting the 32 architectural registers(x0 - x31). The RF is built on top of two FPGA RAM blocks, providing two read ports and one write port, with a single clock cycle latency. The core includes a platform-optimized barrel shifter and multiplier (both with 2 cycle latency). Most of the instruction results are bypassed to achieve interlock-free execution of dependent instructions, either by the Read-After-Writer (RAW) bypass path in the RF or after the X2/W stage. The RISCV Control and Status Registers (CSRs), interrupts and a limited set of exceptions, although the interrupt CSR layout has been simplified to minimize the occupied FPGA area. There is another block to manage the exceptions and interrupts.

IV. Experimental Resources

For this project the author used ARTY A7 board, Vivado 2018.3 and Vivado HLS tools. The ARTY A7, is a ready-to-use development platform designed around the Artix-7

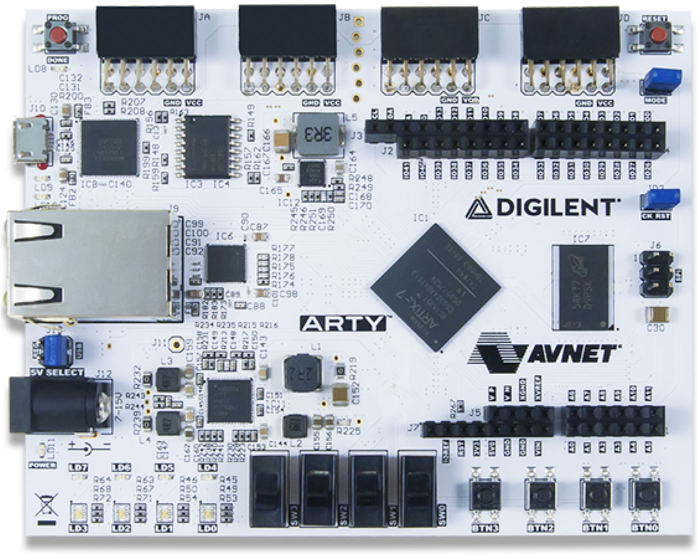


Fig. 2. ARTY A7 Board

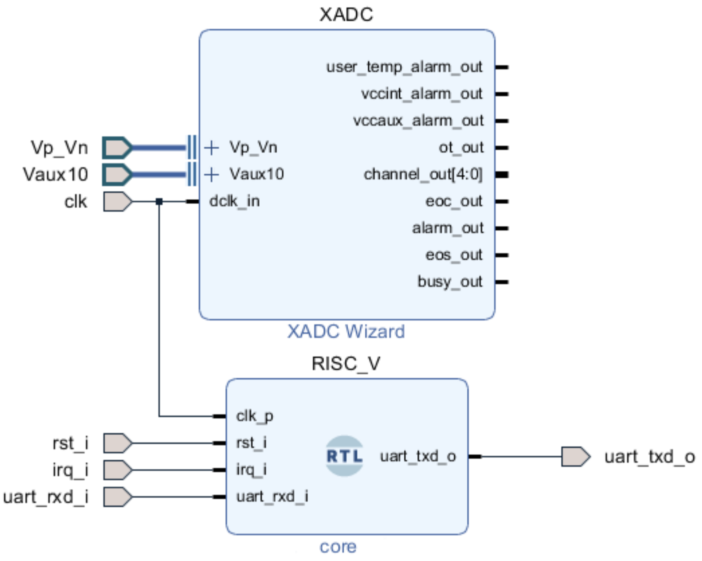


Fig. 3. System Block Diagram

Field Programmable Gate Array (FPGA) from Xilinx (Fig. 2) with features: Artix-7 XC7A35T-L1CSG324I FPGA, On-chip analog-to-digital converter (XADC), Programmable over JTAG and Quad-SPI Flash, 256 MB DDR3L with a 16-bit bus @ 667 MHz, 16 MB Quad-SPI Flash, 10/100 Mb/s Ethernet, USB-UART Bridge, Switches, Buttons, RGB LEDs, Four Pmod interfaces (32 I/O), Arduino/ChipKit “shield” connector (49 I/O).

V. Experimental Work

A. Variable Clock Generator

The hardware system implementation was done using the Block Design facility in Vivado 2018.3. The system block diagram is shown in Fig. 3. It contains an analog-digital converter that allows to determine the current consumption of the entire system. Channel 10 of the converter is connected to an external circuit (INA199A1) of the current measurement consumed by the FPGA. The circuit generates 500mV/A.

RISC V processor presents two external communication pins (uart_rxd/uart_txd) and a pin for interruption (irq). The instruction and data memory is implemented inside of the CPU. All registry blocks have been multiplied: pipeline (F/D, D/Ex, Ex/Wb, Wb/F) and the registers file RF. This allows to retain, at some point, a maximum of four independent threads. The context of each thread is retained in a set of registers. The change of context is made through a switching of registers in a single clock period. Also, each set of registers assigned to a thread is piloted by a programmable clock through scaling registers (CLKSCALE). These registers allow the generation from the system clock of four independent CLK signals and proportionate to the values written in the scaling registers. Each thread will work with a frequency given by its own clock register (Fig. 4).

The multi-pipeline architecture is shown in Fig. 5. It can be seen from the Fig. 1 multiplication of pipeline registers and registers file RF. There is also the clock generation block with the four CLKSCALE registers. CLKSCALE registers were mapped in free space of the CSR registers (Control and Status Registers) to addresses 0x0E00-0x0E03. In Fig. 5, the clock generator block diagram is also observed. Its main components are the registry for dividing the clock and the thread selection unit. When a specific thread is selected, the clock signal assigned to it is also activated. That clock is proportional to the value written in the CLKSCALE register dedicated to the thread.

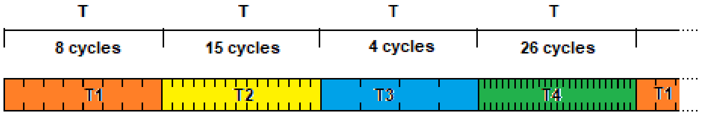


Fig. 4. Schematic Presentation of 4 Threads Running

Listing 1: CLKSCALE registers selection.

```
/**************************************/
    wire wr_genclk;
    assign wr_genclk = (dm_addr_o == 32'h0E00)
    || (dm_addr_o == 32'h0E01)
    || (dm_addr_o == 32'h0E02)
    || (dm_addr_o == 32'h0E03);

    clk_gen clkgen (
          .clock(clkin),
          .reset(1'b0),
          .wr(wr_genclk && dm_store_o),
          .inter(irq_i),
          .data(dm_data_s_o),
          .addr(dm_addr_o[1:0]),
          .selchn(slct),
          .selchnout(select),
          .clkout(clk_i)
);
/**************************************/
```

The Listing 1 also shows the Verilog interface with the clock generation block. This block shows three signals (*wr*, *data*, and *addr*) used to write the four scaling registers. It also presents an interruption signal (*inter*) that announces this block that an interruption occurred. The signals (*selchn* and *selchnout*) are used in the selection of the registers set assigned to each thread. The *clkout* signal is the clock signal that drives the threads.

The way to access CLKSCALE registers in C is done using pointers:

```
  volatile int *CLKSCALE0 = (int*)0x0E00;
*CLKSCALE0 = 0x03ff;
```

The default value of the CLKSCALE registers is 0. That is, the thread assigned to that register runs at full frequency. The clock scaling factor is the value of CLKSCALE + 1. In Listing 2 is presented the C code written for the first thread. The program sends a message through USART and sets the frequency with which this thread will run as CLK_MAX/0x0400. The other source programs for threads 1, 2, and 3 are similar. It differs only the message sent by USART and the value written in CLKSCALE. The program runs in an infinite routine and copies the message from the **hello* address to the **TX_REG* address from where it is sent to USART.

Each thread is called periodically. It will send your own message to USART and run it with its own frequency (Fig. 4).

Listing 2: C code for thread 0.

```
/*************************************/
  const char *hello="Hello␣sCPU0";

  volatile int *TX_REG = (int*)0x0F00;
  volatile int *CLKSCALE0 = (int*)0x0E00;

void main()
{
  char *s = (char*)hello;
  *CLKSCALE0 = 0x03ff;

  while(1){
    s = (char*)hello;
    TX_REG = (int*)0x0F00;
    while(*s) {
        *TX_REG++ = *s++;
    }
  }
}
/*************************************/
```

The result of the four thread execution is shown in Fig. 6 (the most important signals). You can see the thread selection signals, the values written in the CLKSCALE registry, the clock signals and the data sent by USART (Fig. 6a). The writing of the CLKSCALE registers is made when the SELECT_TH signal takes the corresponding value. That means that the code is running for each thread in part: e.g. when the thread 1 is active it writes CLKSCALE1 with the value 0x10, etc.

Fig. 6b shows how to switch the thread 0 with thread 1. You can also see the change of the clock. When the thread changes, it is expected to finish the last clock period in thread 0 and start the first clock period in thread 1. The clock's frequency for thread 1 is 17 times less than the frequency of thread 0 (0x10 + 1).

As in Fig. 6c, switching clocks happens after changing the threads at the end of the last clock period.

Listing 3: ASM code for thread 0. (fragment)

```
/**************************************/
38:     fec42783    lw    a5,-20(s0)
3c:     00178713    addi  a4,a5,1
44:     0007c603    lbu   a2,0(a5)
48:     08802783    lw    a5,136(zero)
4c:     00478693    addi  a3,a5,4
50:     08d02423    sw    a3,136(zero)
54:     00060713    mv    a4,a2
58:     00e7a023    sw    a4,0(a5)
5c:     fe842783    lw    a5,-24(s0)
60:     00178793    addi  a5,a5,1
64:     fef42423    sw    a5,-24(s0)
68:     fec42783    lw    a5,-20(s0)
6c:     0007c783    lbu   a5,0(a5)
70:     fc0794e3    bnez  a5,38 <main+0x38>
74:     fadff06f    j     20 <main+0x20>
/**************************************/
```

In Listing 3 a fragment of the ASM code generated in the compilation is presented. Fig. 6d shows the last instruction in the thread 0 performed when switching to thread 1. When the thread returns to execution 0, it resumes its execution from the point where it was suspended. (red circles)

In Fig. 7, the result of running the four threads on the multi-pipeline RISC V system is presented. Each thread runs at different frequency but sends its message to TX USART.

Considering that interruptions must be taken into consideration immediately after their appearance, the problem that arises when an interruption occurs and the current thread in execution operates at a small frequency, it must accelerate the execution of the last instruction and activate immediately the thread dedicated to that interrupt.

In Fig. 8, the response time of the system to the occurrence of an interruption is presented. It is noted that the interruption signal *irq_i* is activated. Immediately the execution of the last instruction is accelerated in thread 1 (0x6c6...) and proceed to the execution of the interruption handler (in our case the thread 0). The worst case (Fig. 8a), is when the first instruction in the interruption handler is executed after three cycles of the system clock (*clkin*). The most favorable case is shown in Fig. 8b and it appears when the first instruction in the interrupt routine is executed after two cycles.

The Fig. 9 shows the flow chart of the clock generation block, implemented in the verilog and used as IP block. The program checks if the interrupt is set. If so, make the output *clkout=0* and *clkout=1* with *contor=CLKSCALE0* (the default frequency for interrupt routine). Otherwise, depending on the values from CLKSCALE registers, is generate clkout signal.

The code handles the situation when an interruption occurs or when it is inactive. When an interruption occurs, it should be addressed immediately. If, at the time of interruption, the clock signal is on high level then it must be switched to low

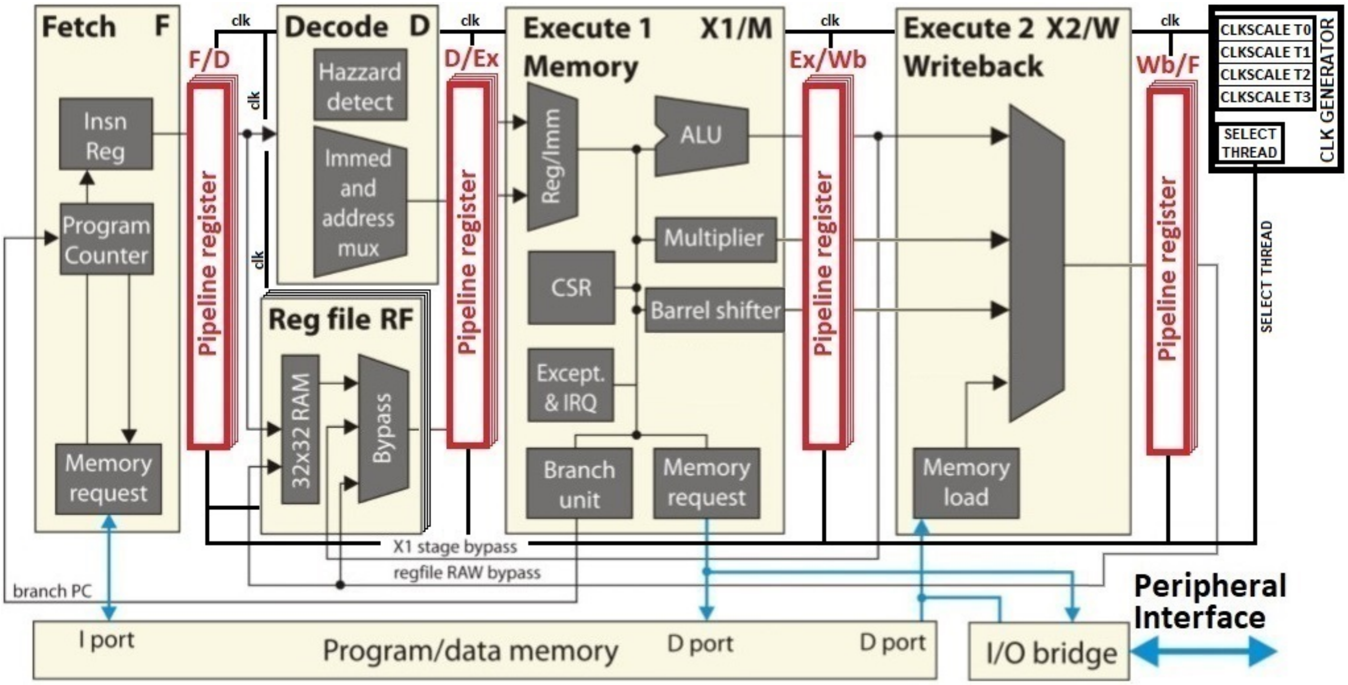


Fig. 5. Multi-Pipeline Architecture with CLKSCALE Registers for each Pipeline Registers Set

level, after which the first cycle of fetching instructions from the interruption handler and the change of frequency are started (Fig. 8a). If, when the interruption occurs, the clock signal is on low level, then at the next cycle the first instruction from the interruption handler is fetched and the frequency is changed (this is the most favorable situation Fig. 8b).

In the situation when no interruption occurs, then when the context changes it waits until the last instruction from the previous thread is executed, and the following execution instructions in the active thread are started when the frequency changes. The frequency is changed based on the counters read from the CLKSCALE registry.

B. Measuring the Energy Consumed

The energy used by the system was measured with the XADC IP block (Fig. 3). Depending on the current consumed by the entire system implemented on the FPGA, the voltage on the XDAC Channel 10 (Vaux10) changes. At each 500 mV measured on Vaux10, the FPGA consumes 1A.

According to equations 1-5 the use of lower frequencies should result in the decrease of the current consumption.

Several measurements have been made with various CLKSCALE values. Initially, measurements of the current consumed with identical values of the four CLKSCALE registers were made, with values ranging from 0x0fff to 0x0000. The first lines in Table I show the current consumed at these CLKSCALE values. The last lines in this table show the current consumed with random values of the CLKSCALE registry.

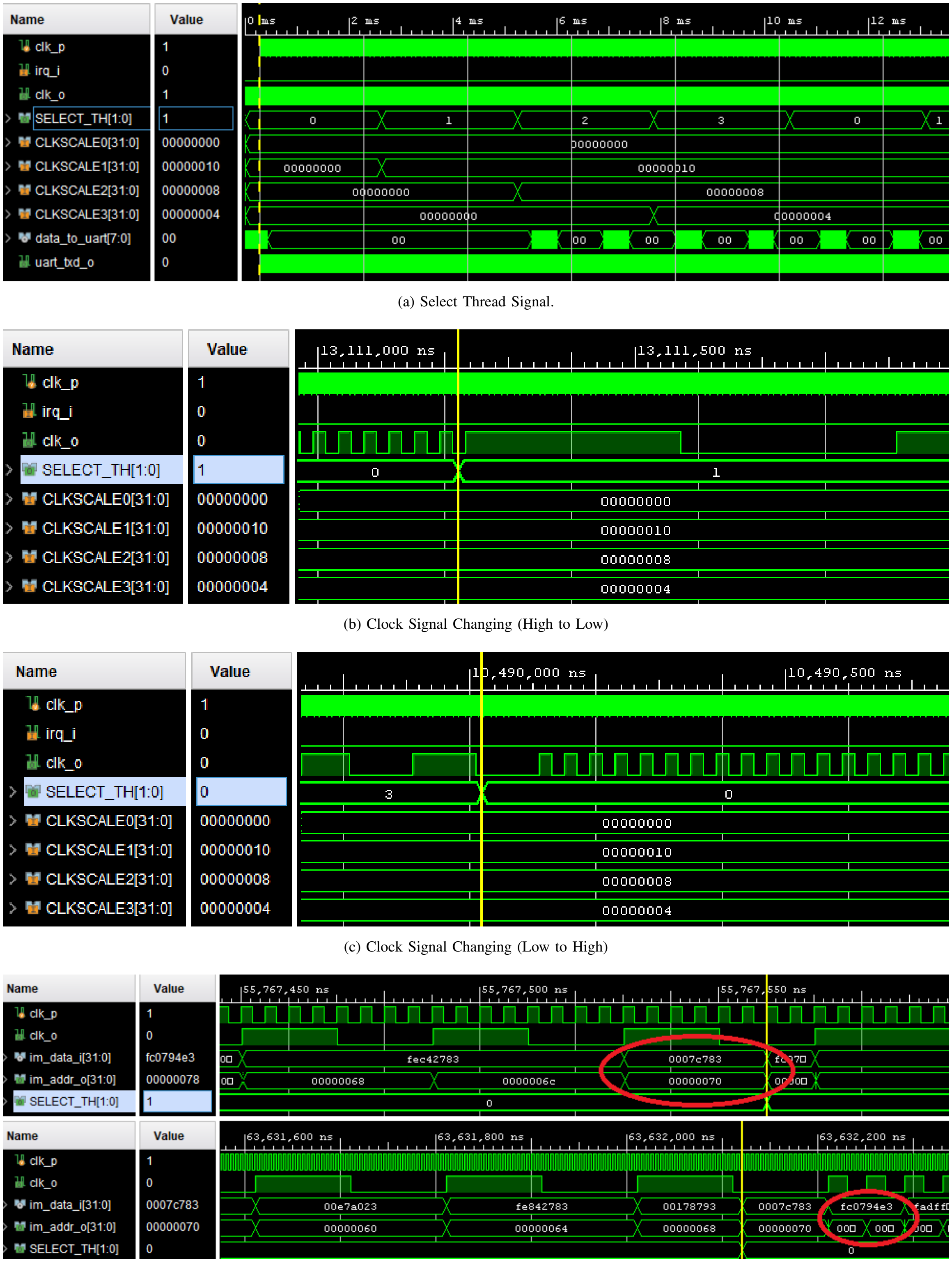
You can see a decrease in the current consumed when the semi-processors operated at lower frequencies. If this does not affect the functioning of the whole system as a whole then we can say that we have achieved a reduction of the energy consumed without low performance. From the Table I you can see that between the maximum current and the minimum current consumed it is a ratio of about 30%. If it is also taken into account the current consumed by the XADC circuit that has been implemented only to perform measurements, then we can talk about a yield greater than 30% of saved energy.

TABLE I. CURRENT CONSUMED ACCORDING TO THE VALUES WRITTEN IN CLKSCALE REGISTERS

Vaux10	A	CLKSC0	CLKSC1	CLKSC2	CLKSC3
0.012V	24mA	0x0fff	0x0fff	0x0fff	0x0fff
0.013V	26mA	0x3ff	0x3ff	0x3ff	0x3ff
0.014V	28mA	0x0ff	0x0ff	0x0ff	0x0ff
0.014V	28mA	0x00f	0x00f	0x00f	0x00f
0.015V	30mA	0x003	0x003	0x003	0x003
0.016V	32mA	0x002	0x002	0x002	0x002
0.017V	34mA	0x001	0x001	0x001	0x001
0.018V	36mA	0x000	0x000	0x000	0x000
0.014V	28mA	0x010	0x0f0	0x010	0x001
0.014V	28mA	0x000	0x010	0x100	0x003
0.016V	32mA	0x000	0x004	0x003	0x000
0.013V	23mA	0x000	0xfff	0x0ff	0xfff

VI. CONCLUSION

The conclusions drawn from this study are as follows: by multiplying the pipeline registers and the registry file of a RISC V architecture, four semi-processors using the same hardware resources have been obtained. Each semi-processor runs a thread at different frequencies. Switching between threads is done in a clock cycle due to the multiplication of pipeline



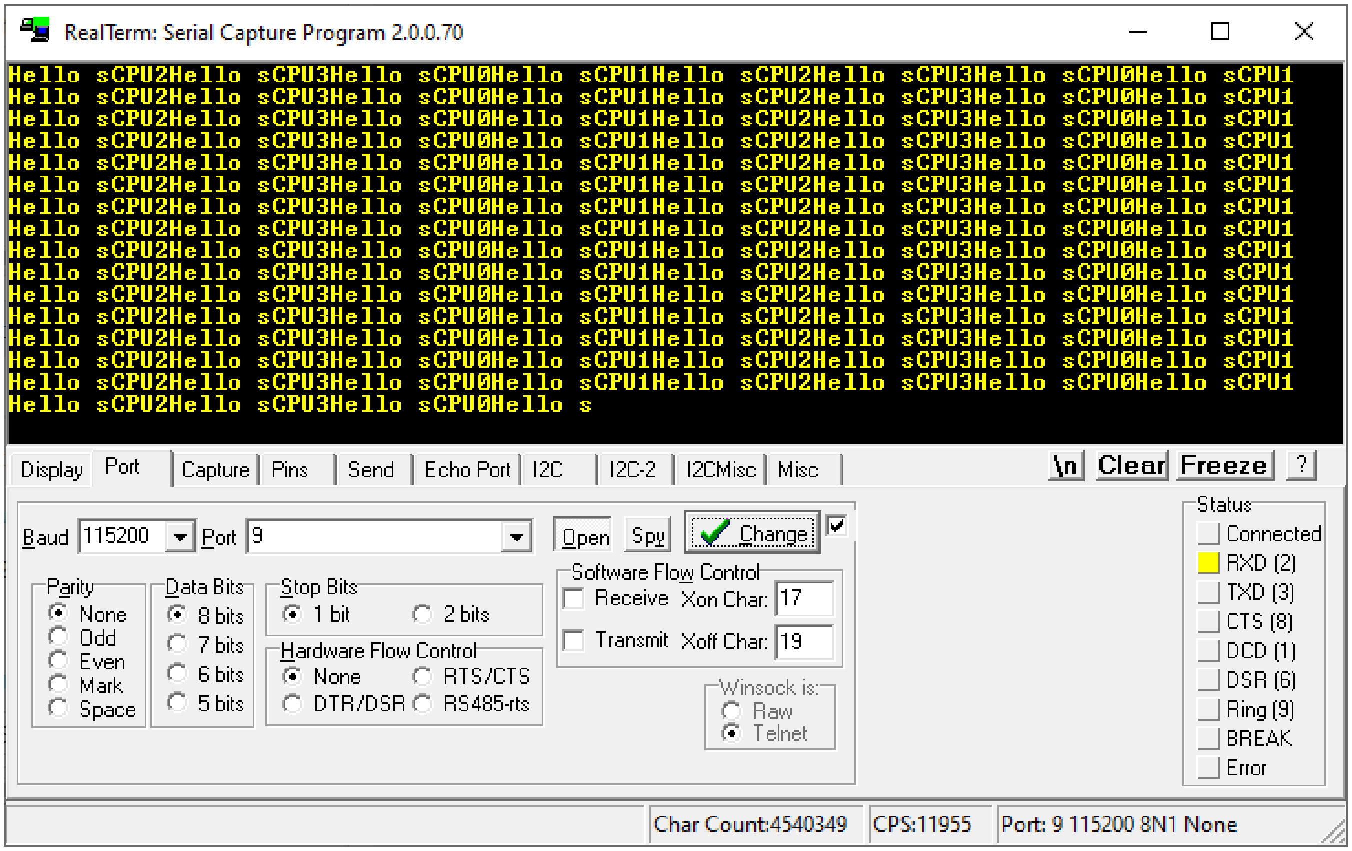
(a) Select Thread Signal.

(b) Clock Signal Changing (High to Low)

(c) Clock Signal Changing (Low to High)

(d) Last and Next Instruction in Thread 0.

Fig. 6. The Most Important Signals.



RealTerm: Serial Capture Program 2.0.0.70

```
Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1
Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1
Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1
Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1
Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1
Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1
Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1
Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1
Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1
Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1
Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1
Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1
Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1
Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1
Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1Hello sCPU2Hello sCPU3Hello sCPU0Hello sCPU1
Hello sCPU2Hello sCPU3Hello sCPU0Hello s
```

Baud 115200 Port 9

Parity: ● None ○ Odd ○ Even ○ Mark ○ Space

Data Bits: ● 8 bits ○ 7 bits ○ 6 bits ○ 5 bits

Stop Bits: ● 1 bit ○ 2 bits

Hardware Flow Control: ● None ○ RTS/CTS ○ DTR/DSR ○ RS485-rts

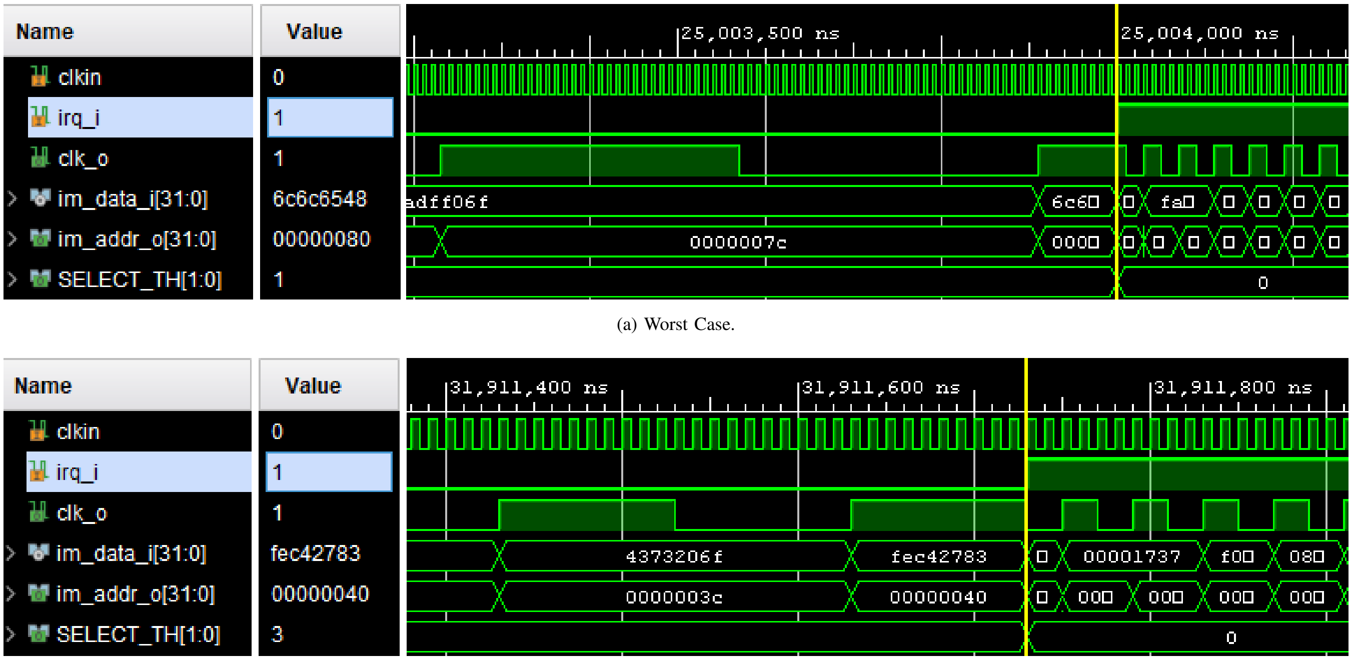
Software Flow Control: □ Receive Xon Char: 17 □ Transmit Xoff Char: 19

Winsock is: ○ Raw ● Telnet

Status: □ Connected 🟨 RXD (2) □ TXD (3) □ CTS (8) □ DCD (1) □ DSR (6) □ Ring (9) □ BREAK □ Error

Char Count:4540349 CPS:11955 Port: 9 115200 8N1 None

Fig. 7. RealTerm Terminal with Messages from all Threads



(a) Worst Case.

(b) Best Case.

Fig. 8. Interrupt Response Time.

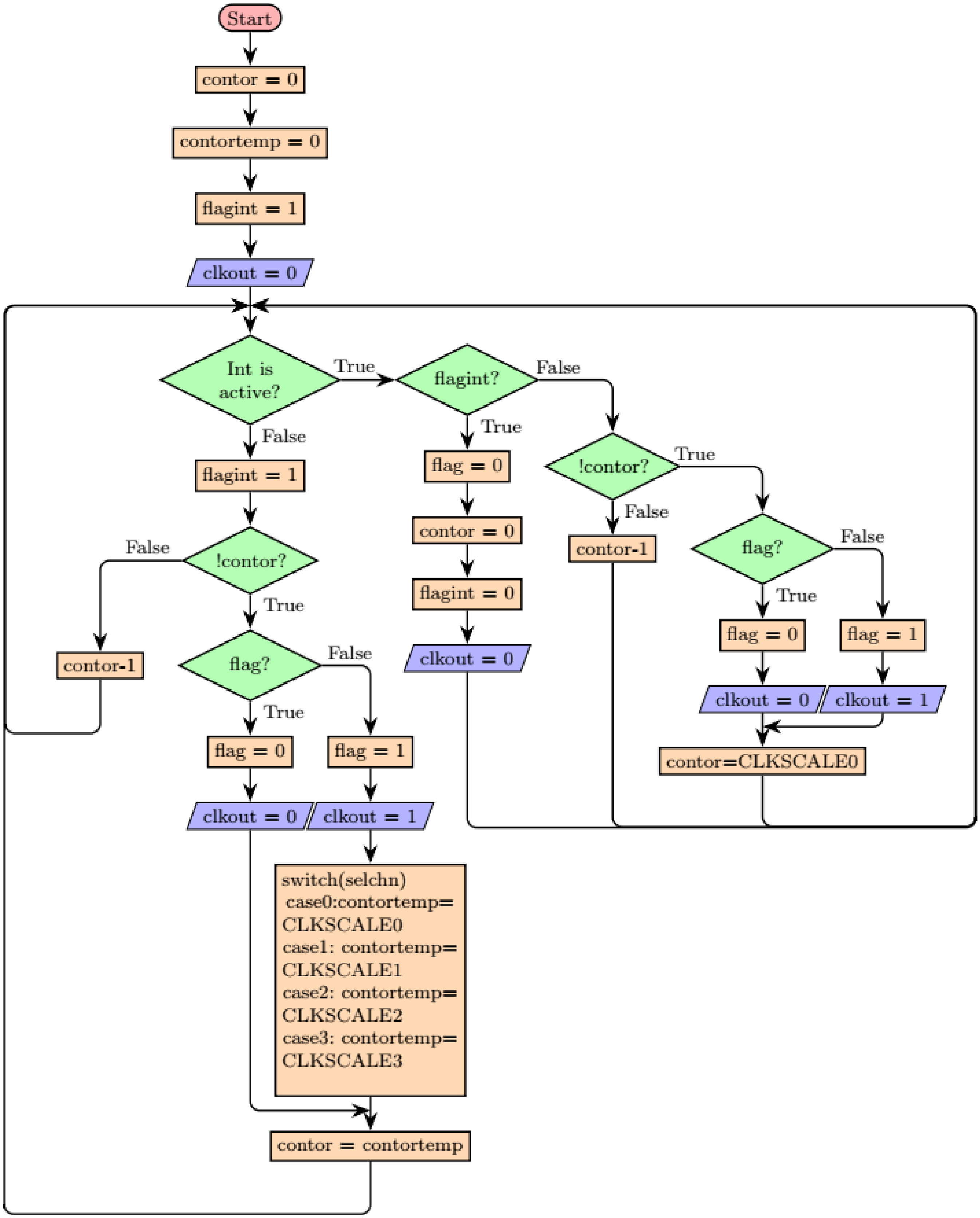


Fig. 9. Flow Chart of Clock Generator Code

registers. Depending on the needs of real-time responses of each thread, the Hard Real Time threads will work at high frequencies and the Soft Real Time threads will work at lower frequencies. In this way, a lower energy consumption will be achieved due to the fact that the energy consumed is proportional to the system's working frequency.

As future work the author wants to create an auto-tuning architecture adapted to the priorities and response time required for each task. If a task has a much shorter response time than a response time that does not generate errors in the operation of the system, it will decrease its execution frequency until it reaches close to this time without exceeding it. It will execute the task in real time with a minimum energy consumption. Thus the author aims to implement a block that detects these malfunctions due to an inadequate response time.

Acknowledgment

The infrastructure used for this work was partially supported by the project "Integrated Center for research, development and innovation in Advanced Materials, Nanotechnologies, and Distributed Systems for fabrication and control", Contract No. 671/09.04.2015, Sectoral Operational Program for Increase of the Economic Competitiveness co-funded from the European Regional Development Fund.

References

[1] Ian Finlayson, Gang-Ryung Uh, David Whalley and Gary Tyson, *Improving Low Power Processor Eficiency with Static Pipelining*, 15th Workshop on Interaction between Compilers and Computer Architectures, 2011.

[2] S. Borkar, *Design challenges of technology scaling*, In IEEE Micro, Vol. 19, Issue 4, pp. 23-29, 1999.

[3] M. Broersma, *Intel chip not ready for the cool crowd*, CNET Tech News, http://news.com.com/ 2100-1001_3-271443.html

[4] D. Brooks et. al., *Wattch: A Framework for Architectural-level Power Analysis and Optimizations*, In Proceedings of the 27th International Symposium on Computer Architecture, pp. 83-94, 2000.

[5] M. Edahiro et.al., *A single-chip multiprocessor for smart terminals*, In IEEE Micro, Volume 20, Issue 4, pp. 12-20, July 2000.

[6] M. Fleischmann, *Longrun Power Management*, Transmeta Corporation, January 2001.

[7] R. Fromm et al., *The Energy Efficiency Of Iram Architectures*, In Proceedings of IEEE International Symposium on Computer Architecture, pp. 327-337, 1997.

[8] L. Geppert, T. S. Perry, *Transmeta's magic show*, In IEEE Spectrum, Vol. 37, Issue 5, pp. 26-33, 2000.

[9] R. Gonzalez, M. Horowitz, *Energy dissipation in general purpose processors*, In IEEE Journal of Solid-State Circuits, Volume 31, Issue 9, pp. 1277-1284, September 1996.

[10] K. Flautner, N. S. Kim, S. Martin, D. Blaauwm, T. Mudge, *Drowsy Caches: Simple Techniques for Reducing Leakage Power*, In Proceedings of the International Symposium on Computer Architecture (ISCA-29), Anchorage, Alaska, 2002.

[11] T. R. Halfhill, *Transmeta breaks x86 low power barrier*, Microprocessor Report, pp.9-18, February 2000.

[12] N. S. Kim et. al., *Leakage Current: Moore's Law Meets Static Power*, In IEEE Computer, Volume 36, Issue 12, pp. 68-75, 2003.

[13] A. Gomez1, C. Pinto, A. Bartolini et. al., *Reducing Energy Consumption in Microcontroller-based Platforms with Low Design Margin Co-Processors*, in Proc. DATE, 2015.

[14] R. Kumar, K. Farkas, N. P. Jouppi, P. Ranganathan, and D. M. Tullsen, *Processor power reduction via single-ISA heterogeneous multi-core architectures*, In IEEE Computer Architecture Letters, vol. 2, no. 1, 2003.

[15] U. R. Karpuzcu, A. Sinkar, N. S. Kim, and J. Torrellas, *EnergySmart: Toward Energy-efficient Manycores for Near-Threshold Computing*, In ACM HPCA, 2013.

[16] R. Marau, P. Leite, and M. Velasco, *Performing flexible control on lowcost microcontrollers using a minimal real-time kernel*, In IEEE Trans. Ind. Informat, vol. 4, no. 2, 2008.

[17] T.Włostowski,J.Serrano, *Developing Distributed Hard-Real Time Software Systems Using FPGAs and Soft Cores*, Proceedings of ICALEPCS2015, Melbourne, Australia-Pre-Press Release23-Oct-2015.

Cryptanalysis and Countermeasure of "An Efficient NTRU-based Authentication Protocol in IoT Environment"

YoHan Park[1], Woojin Seok[2], Wonhyuk Lee[3], Hong Taek Ju[4]
Department of Computer Engineering
Keimyung University, Korea, Republic[1,4]
KREONET Center, KISTI, Korea, Republic[2,3]

***Abstract*—A quantum computer is a paradigm of information processing that can show remarkable possibilities of exponentially improved information processing. However, this paradigm could disrupt the current cryptosystem, which is called quantum computing attacks, by calculating factoring problem and discrete logarithm problem. Recently, NTRU is applied to various security systems, because it provides security against to provide security against quantum computing attacks. Furthermore, NTRU provides similar security level and efficient computation time of encryption/decryption compared to traditional PKC. In 2018, Jeong et al. proposed an user authentication and key distribution scheme using NTRU. They claimed that their scheme provides various security properties and secure against quantum computing attacks. In this paper, we demonstrate that their scheme has security pitfalls and incorrectness in login and authentication phase. We also suggest countermeasures to fix the incorrectness and provide security against various attacks.**

Keywords*—*Post-quantum; NTRU; biometrics; user authentication; key agreement

I. Introduction

Development of Internet of Things (IoT) technology help users connect service providers easily and fast and utilize various services such as Health care, SmartHome, SmartGrid, and so on. However, IoT environments have threats to security and privacy because of its wireless nature [1]. Such threats hinder users make use of beneficial applications and service providers may not continue to invest for profits. Security problems should get solved to make the IoT-based services widely spread and applied.

User authentication and key agreement are essential requirement among all other security concerns. Those security services provide integrity and confidentiality for IoT environments [2]–[4]. Malicious adversaries will freely access user's critical and valuable information if secure authentication and key agreement methods do not provide.

Security of public key cryptosystem (PKC) are mostly based on the difficulty of factorization problem (FP) or discrete logarithm problem (DLP). RSA and elliptic curve cryptosystem (ECC) are major the examples of current cryptosystems based on FP and DLP. However, these algorithms are vulnerable to a quantum computing attack. In 1994, Peter W. Shor [5] proposed a quantum computing algorithm which can solve FP efficiently. And a quantum search algorithm proposed by Grover [6] can easily solve DLP. These algorithms based on quantum computing became major threats to all security protocols using RSA or DLP. Therefore, ETSI [7] and NIST [8] recommended that post-quantum cryptosystem (PQC) should be prepared with high priority.

There are several PQC which are secure against quantum computing attacks. These cryptosystems use Code, Lattice, Hash and Multivariate to provide security in quantum computing environments. Among many algorithms of PQC, NTRU, proposed by J. Hoffstein [9] in 1996, has been approved for standardization by Institute of Electrical and Electronics Engineerings (IEEE) [10]. The security of NTRU is based on the difficulty on a finding shortest path in n-th degree Lattice. Comparing to traditional PKC, NTRU provides not only similar security level, but also efficient computation time of encryption/decryption. Recently, NTRU is applied to various security systems which provides user authentication and key agreement.

Recently, a number of authentication and key agreement schemes have been proposed in IoT environments [11], [12], [19]. In 2017, Li et al. [13] proposed a key distribution protocol using ECC in IoT. However, the scheme is vulnerable to quantum computing attacks, such as Shor [5] or Grover [6] algorithm. To overcome these security pitfalls, Jeong et al. proposed an efficient NTRU-based authentication protocol in IoT environments [14] in 2018. They proposed user authentication and key agreement protocols using NTRU and claimed their scheme is secure against quantum computing attacks and prevents impersonations attack and session key disclosure attack. However, we find out that their scheme does not provide a proper user authentication process and is weak to various attacks, such as privileged insider attacks, impersonation attacks, and session key disclosure attacks. In addition, we show their scheme fails to provide correctness in login phase and authentication phase.

In this paper, we show the security weaknesses of Jeong et al.'s scheme. In addition, we propose countermeasures for the weaknesses of Jeong et al.'s scheme.

A. Contributions

The contributions made in the paper are listed below:

1) We demonstrate that Jeong et al.'s scheme has an incorrectness in login phase and authentication phase.
2) We analyze security weaknesses of Jeong et al.'s scheme and show that their scheme is vulnerable to privilege insider attacks, impersonation attacks, and session key disclosure attacks.
3) We propose countermeasures to overcome the security weaknesses of Jeong et al.'s scheme. The countermeasures help to prevent various attacks such as password guessing attacks, user impersonation attacks and session key disclosure attacks from malicious adversaries.

B. Paper Structure

The rest of the paper is organized as follows. In Section II, we introduce preliminaries used in this paper. In Section III, we review Jeong et al.'s scheme followed by the cryptanalysis of Jeong et al.'s scheme in Section IV. In Section V, we propose countermeasures for the weaknesses of Jeong et al.'s scheme. Finally, Section VI concludes the paper.

II. Preliminaries

A. NTRU

NTRU is a lattice-based public key cryptosystem proposed by Jeffry Hoffstein et al. [9]. This provides a similar security level, but high performance compared to RSA and ECC because of low computational complexity of polynomial convolution operation. NTRU requires $O(n^2)$ operations to encrypt or decrypt a message of size n, but RSA and ECC require $O(n^3)$ operations. NTRU, furthermore, resists quantum computing attacks and has adopted standard as IEEE 1363.1 and X9.98. NTRU cryptosystem consists of three parts: key generation, encryption and decryption.

1) Key generation: Alice and Bob are required to generate private/public key in advance to exchange data securely in PKC. The detail steps of key generation are as follows:

Step 1: Alice chooses two polynomials f and g with degree $N-1$ and coefficients in $\{-1,0,1\}$.

Step 2: The polynomial $f \in L_f$ must have the inverse element for modulo p and q. Alice computes $f * f_p^{-1} \equiv 1 \pmod{p}$ and $f * f_q^{-1} \equiv 1 \pmod{q}$.

Step 3: If f does not have an inverse element, Alice turns back to Step 1 and chooses another f. Otherwise, Alice computes the public key $h = pf_q^{-1} * g \pmod{q}$.

f and g are private keys and h is a public key of Alice.

2) Encryption: If Bob wants to send a message to Alice securely, Bob performs the encryption as follows:

Step 1: Bob who wants to send a plaintext polynomial $m \in L_m$ chooses a random polynomial $r \in L_r$ with $N-1$ degree and small coefficients. Coefficients are not restricted to the set $\{-1,0,1\}$.

Step 2: Bob encrypts the message m into e using the public key h of Alice. $e = r * h + m$

Step 3: Bob sends the encrypted message e to Alice.

3) Decryption: After receiving e from Bob, Alice decrypts the message as follows:

Step 1: Alice calculates a convolution $a = e * f \pmod{q}$, where f is a private key of Alice. The coefficient of a should satisfy $A \leq a_i \leq A + q$.

Step 2: Alice retrieves $m \equiv a \pmod{p}$.

B. Notations

Table I describe the notations used throughout the paper.

TABLE I. Notations

Notation	Meaning
U_A	user A
ID_A	identity of U_A
PW_A	password of U_A
RPW_A	pseudo password of U_A
B_A	biometric template of U_A
SC_A	smart card of user U_A
GWN	gateway node
$*$	convolution computation
f, g	private key polynomial $f \in L_f, g \in L_g$
f_p^{-1}, f_q^{-1}	inverse polynomial of f
h	public key
H	hash function
$\|\|$	concatenate operation
$\bigoplus$	XOR operation

III. Review of Jeong et al.'s Scheme

In this section, we review Jeong et al.'s NTRU-based authentication scheme. The scheme is composed of three phases: user registration phase, things registration phase, and login-authentication-key distribution phase.

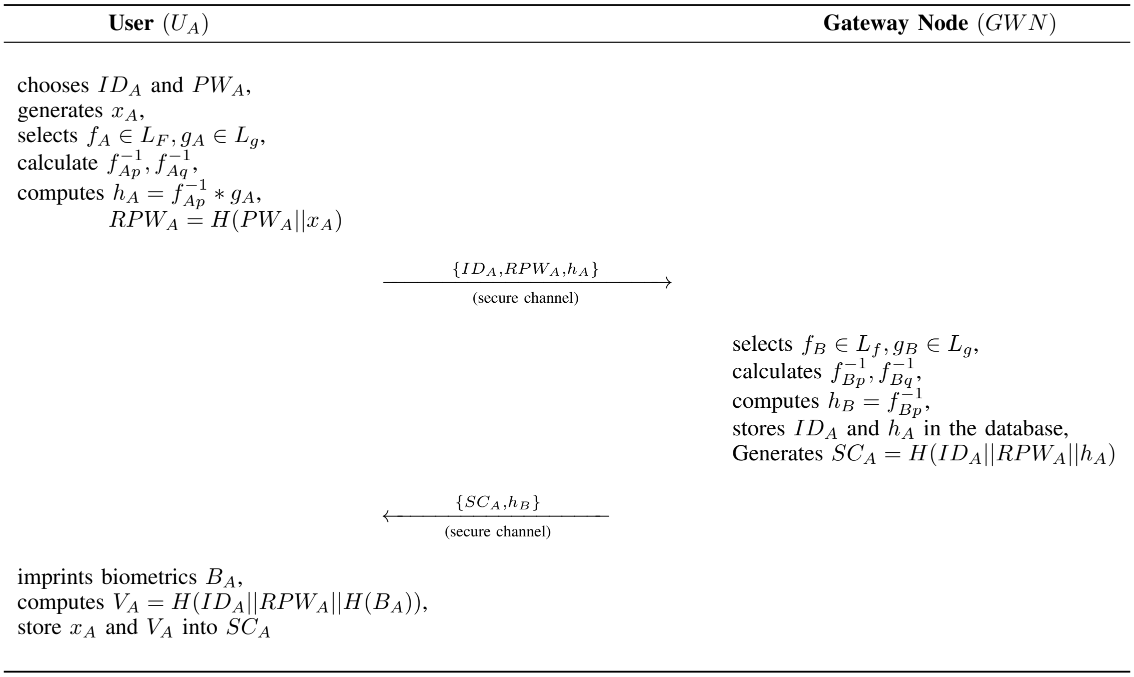
A. User Registration Phase

In this phase, a user registers his/her information to the gateway node, and acquires a personalized smart card SC_A. The Jeong et al.'s user registration phase is illustrated in Fig. 1, and the detailed steps of this registration phase are as follows:

Step 1: A user U_A chooses ID_A and PW_A, then generates a random number x_A. Then, U_A selects polynomials $f_A \in L_f$ and $g_A \in L_g$, then calculates inverse elements f_{Ap}^{-1} and f_{Aq}^{-1} of f. Next U_A calculates the public key $h_A = pf_{Ap}^{-1} * g_A(modq)$ and the pseudo password $RPW_A = H(PW_A||x_A)$. U_A sends the registration request message $\{ID_A, RPW_A, h_A\}$ to the gateway node via a secure channel.

Step 2: After receiving registration request message from the user, the gateway node GWN stores the pair $\{ID_A, h_A\}$ in database. GWN also selects polynomials $f_B \in L_f$ and $g_B \in L_g$, then calculates inverse elements f_{Bp}^{-1} and f_{Bq}^{-1} of f. Next GWN calculates the public key $h_B = pf_{Bp}^{-1} * g_B(modq)$. Next, GWN computes $H(ID_A||RPW_A||h_A)$, and issues a smart card SC_A with $H(ID_A||RPW_A||h_A)$ and sends $\{SC_A, h_B\}$ to U_A.

Step 3: After receiving $\{SC_A, h_B\}$ from the GWN, U_A computes $V_A = H(ID_A||RPW_A||H(B_A))$, where B_A



User (U_A) | **Gateway Node** (GWN)

chooses ID_A and PW_A,
generates x_A,
selects $f_A \in L_F, g_A \in L_g$,
calculate f_{Ap}^{-1}, f_{Aq}^{-1},
computes $h_A = f_{Ap}^{-1} * g_A$,
$RPW_A = H(PW_A||x_A)$

$$\xrightarrow[\text{(secure channel)}]{\{ID_A, RPW_A, h_A\}}$$

selects $f_B \in L_f, g_B \in L_g$,
calculates f_{Bp}^{-1}, f_{Bq}^{-1},
computes $h_B = f_{Bp}^{-1}$,
stores ID_A and h_A in the database,
Generates $SC_A = H(ID_A||RPW_A||h_A)$

$$\xleftarrow[\text{(secure channel)}]{\{SC_A, h_B\}}$$

imprints biometrics B_A,
computes $V_A = H(ID_A||RPW_A||H(B_A))$,
store x_A and V_A into SC_A

Fig. 1. User Registration Phase of Jeong et al.'s Scheme

is the biometric information of U_A. Then U_A stores V_A and x_A in SC_A.

B. Things Registration Phase

In this phase, a thing registers its information to the gateway node GWN, and receives an ephemeral key z_S from GWN. The Jeong et al.'s things registration phase is illustrated in Fig. 2, and the detailed steps of this registration phase are as follows:

Step 1: A thing chooses a random number n_S, then sends it to GWN through the secure channel.

Step 2: After receiving n_S from the thing, GWN choose a random number n_Z. Then GWN stores the pair (n_S, n_Z) in the database. Finally, GWN sends z_S to the thing via secure channel.

Step 3: After receiving n_Z from GWN, the thing stores n_Z in it's database.

C. Login-Authentication-Key Distribution Phase

In this phase, a user uses his/her multi-factor keys to login and authenticate oneself with GWN. Then a user shares session key SK with a thing. The Jeong et al.'s login-authentication-key distribution phase is illustrated in Fig. 3, and the detailed steps of this registration phase are as follows:

Step 1: A user inputs ID_A, PW_A, and imprints the biometrics B_A into the smart card SC_A. Then, SC_A computes RPW_A, V_A', and verifies the validity of the user as follows:

$$\begin{aligned} RPW_A &= H(PW_A||x_A) \\ V_A' &= H(ID_A||RPW_A||H(B_A)) \\ \text{verifies } V_A' &\overset{?}{=} V_A \end{aligned}$$

If it is wrong, SC_A quits the login process. Otherwise, SC_A chooses random numbers r_A, k_A and computes I_A, e_A as follows:

$$\begin{aligned} I_A &= H(ID_A||RPW_A) \\ e_A &= pr_A * h_B + k_A \end{aligned}$$

Then, the user sends $\{I_A, e_A, h_A\}$ to GWN.

Step 2: After receiving $\{I_A, e_A, h_A\}$, GWN verifies I_A using the stored pair $\{ID_A, h_A\}$. Then, GWN retrieves k_A, and computes c_B as follows:

$$\begin{aligned} I_A &= H(ID_A||RPW_A) \\ a_B &= f_B * e_A (\text{mod } q) \\ k_A &= f_B^{-1} * a_A (\text{mod } p) \\ c_B &= z_S \oplus k_A \end{aligned}$$

Then, GWN sends c_B to the thing.

Step 3: After receiving c_B from GWN, the thing retrieves $k_A = c_B \oplus z_S$. Then the thing chooses a random number k_S and computes a session key $SK = H(k_A||k_S||n_S)$. The thing computes $c_S = k_S \oplus z_S$ and sends it to GWN.

Step 4: After receiving c_S, GWN chooses a random number $r_B \in L_r$ and computes k_S and e_B as follows:

$$\begin{aligned} k_S &= c_S \oplus z_S \\ e_B &= pr_B * h_A + (k_S||n_S) (\text{mod } q) \end{aligned}$$

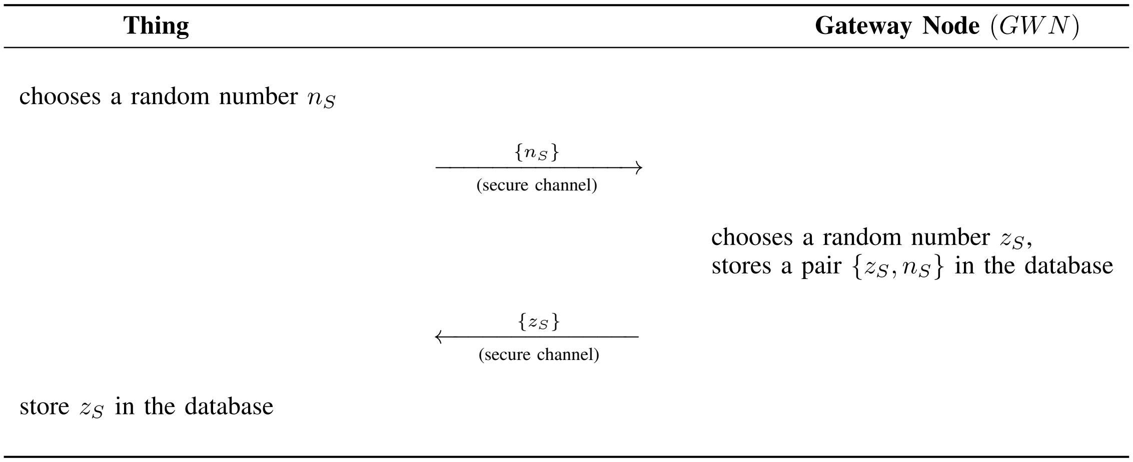

Thing
Gateway Node (GWN)
chooses a random number n_S
$\{n_S\}$
(secure channel)
chooses a random number z_S,
stores a pair $\{z_S, n_S\}$ in the database
$\{z_S\}$
(secure channel)
store z_S in the database

Fig. 2. Thing registration phase of Jeong et al.'s scheme

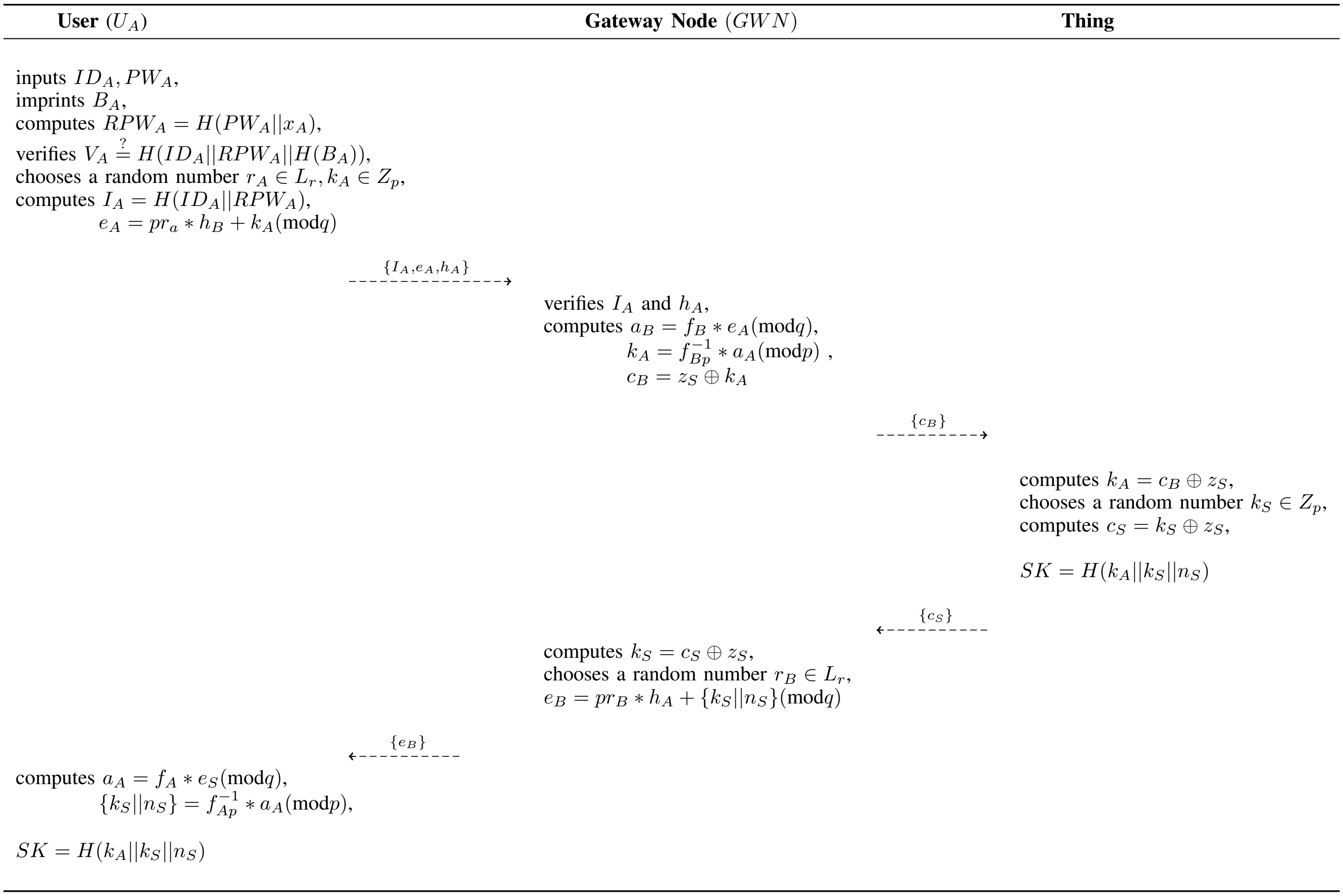

User (U_A)
Gateway Node (GWN)
Thing
inputs ID_A, PW_A,
imprints B_A,
computes $RPW_A = H(PW_A||x_A)$,
verifies $V_A \overset{?}{=} H(ID_A||RPW_A||H(B_A))$,
chooses a random number $r_A \in L_r, k_A \in Z_p$,
computes $I_A = H(ID_A||RPW_A)$,
$e_A = pr_a * h_B + k_A(\text{mod}q)$
$\{I_A, e_A, h_A\}$
verifies I_A and h_A,
computes $a_B = f_B * e_A(\text{mod}q)$,
$k_A = f_{Bp}^{-1} * a_A(\text{mod}p)$,
$c_B = z_S \oplus k_A$
$\{c_B\}$
computes $k_A = c_B \oplus z_S$,
chooses a random number $k_S \in Z_p$,
computes $c_S = k_S \oplus z_S$,
$SK = H(k_A||k_S||n_S)$
$\{c_S\}$
computes $k_S = c_S \oplus z_S$,
chooses a random number $r_B \in L_r$,
$e_B = pr_B * h_A + \{k_S||n_S\}(\text{mod}q)$
$\{e_B\}$
computes $a_A = f_A * e_S(\text{mod}q)$,
$\{k_S||n_S\} = f_{Ap}^{-1} * a_A(\text{mod}p)$,
$SK = H(k_A||k_S||n_S)$

Fig. 3. Login-Authentication-Key Distribution Phase of Jeong et al.'s Scheme

Then, GWN sends e_B to the user.

Step 5: After receiving e_B, the user computes $k_S||n_S$ and the session key SK as follows:

$$\begin{aligned} a_A &= f_A * e_B \\ k_S||n_S &= f_{AP}^{-1} * a_A \\ SK &= H(k_A||k_S||n_S) \end{aligned}$$

IV. Cryptanalysis of Jeong et al.'s Scheme

In this section, we demonstrate the security flaws of Jeong et al.'s scheme. Their scheme does not provide correctness at the login phase and the key distribution phase. Thus, the user who tries to connect a IoT device cannot login to GWN and share a session key. In addition, the scheme is vulnerable to privileged insider attacks. The insider adversaries can sneak into the database of GWN and illegally capture the stored information. The adversaries can guess identity and password of users using the captured information and impersonate a legitimate user.

In this paper, we assumed that an adversary $\mathcal{A}$ could steal or obtain the user's smart card SC_A. In addition, an adversary $\mathcal{A}$ could extract information $\{H(ID_A||RPW_A||h_A), V_A, x_A\}$ from the smart card [15] and could get previous session messages transmitted through public network. The description of the security weaknesses of Jeong et al.'s scheme is as follows.

A. Incorrectness

1) Incorrectness at the login phase: In the login phase, GWN verifies the validity of a user by comparing the received value I_A with the computed value $H(ID_A||RPW_A)$. If it is correct, GWN authenticates the user and proceeds the key distribution phase. GWN can find ID_A using the pair $\{ID_A, h_A\}$ stored in the database. However, GWN cannot get RPW_A in the database and other transmitted values. Therefore, GWN cannot check the legitimacy of a user who wants to access things.

2) Incorrectness at the key distribution phase: To establish a session key between a user and a thing, each party should know the information $\{k_A, k_S, n_S\}$. k_A is a random number generated by a user, and encrypted with the pubic key h_B of GWN. To retrieve k_A from the encrypted message e_A, Jeong et al.'s present the mathematical equation as follows:

$$\begin{aligned} a_B &= f_B * e_A (\text{mod } q) \\ k_A &= f_B^{-1} * a_A (\text{mod } p) \end{aligned}$$

Unfortunately, it is incorrect and cannot find k_A. The equation should be presented as follows:

$$\begin{aligned} a_B &= e_A * f_B (\text{mod } q) \\ &= (pr_A * h_B + k_A) * f_B (\text{mod } q) \\ &= (pr_A * h_B * f_B) + (k_A * f_B) (\text{mod } q) \\ &= k_A * f_B (\text{mod } q) \\ a_B * f_B^{-1} (\text{mod } p) &= k_A * f_B * f_B^{-1} (\text{mod } p) \\ &= k_A (\text{mod } q) \end{aligned}$$

In addition, a user cannot retrieve $\{k_S||n_S\}$ from a_A, because GWN sends e_B but a user decrypt e_S. e_S should be replaced with e_B.

B. Privileged Insider Attack

Jeong et al.'s analyzed their scheme and insisted that the scheme is secure against privileged insider attacks. However, we cryptanalyze and show that their scheme is vulnerable to the attacks. A malicious inside adversary can access to the database and compute user's information, then guess identities of users. Using the information, the adversary can act as a legal user as follows:

1) An insider adversary $\mathcal{A}$ can get the values $\{ID_A, h_A\}$ stored in the database and $\{H(ID_A||RPW_A||h_A), V_A, x_A\}$ from the smart card SC_A.
2) $\mathcal{A}$ performs an offline password guessing attack. $\mathcal{A}$ guesses a password PW'_A and computes $RPW'_A = H(PW'_A||x_A)$.
3) $\mathcal{A}$ compares the computed value $H(ID_A||RPW'_A||h_A)$ with $H(ID_A||RPW_A||h_A)$ which is stored in SC_A. If it matches, $\mathcal{A}$ successfully guesses the password of the user.

Therefore, Jeong et al.'s scheme does not provide security against privileged insider attacks.

C. Impersonation Attack

Jeong et al. claimed that their scheme is secure against impersonation attacks. However, once the inside adversary $\mathcal{A}$ correctly guesses PW_A and finds secret values, such as ID_A and RPW_A, $\mathcal{A}$ can generate the login message $\{I_A, e_a, h_A\}$. Then, $\mathcal{A}$ can impersonate the user.

D. Session Key Disclosure Attack

Jeong et al.'s insisted that the scheme provides session key disclosure attacks. But, we show that their scheme is weak to this attack. An inside adversary can access to the database and obtain secret information pair of Things $\{n_S, z_S\}$. The adversary can compute a session key using the information as follows:

1) The insider adversary $\mathcal{A}$ who knows $\{n_S, z_S\}$ can acquire c_B and c_S which are transferred via a insecure channel.
2) $\mathcal{A}$ can compute $k_A = c_B \oplus z_S, k_S = c_S \oplus z_S$, because $\mathcal{A}$ knows z_S from the database and c_B and c_S from an insecure channel.
3) $\mathcal{A}$ who successfully computes k_A and k_S can finally derive a session key $SK = H(k_A||k_S||n_S)$, because n_S is also disclosed.

Therefore, Jeong et al.'s scheme is vulnerable to session key disclosure attacks.

V. Countermeasures

In this section, we present the fixes for the incorrectness and the countermeasures to improve the security weakness of Jeong et al.'s scheme.

A. Fixes for the Incorrectness

Jeong et al.'s scheme cannot provide the user authentication, because GWN cannot computes I_A using the data in the database. GWN should know the pair $\{ID_A, RPW_A\}$ to compute I_A and verify the validity of user U_A. Therefore, GWN should store the three sets $\{ID_A, RPW_A, h_A\}$ in the database at the user registration phase.

Unfortunately, if GWN store that tuple in the database, privileged insider can easily obtain ID_A and RPW_A, and compute I_A without password guessing process. To solve this problem, I recommend not to allow storing the identity and password of a user in the database. Instead, a user generates a pseudo-identity and sends it to GWN for the verification. There are many authentication schemes which do not allow to store the identity of a user but provide an authentication of a user [16]–[18].

The fixes for the incorrectness at the key distribution phase are introduced at Section IV.

B. Countermeasure of Privileged Insider Attacks

The privileged insider adversary $\mathcal{A}$ can use the data ID_A stored in the database. $\mathcal{A}$ uses this identity and the data $H(ID_A||RPW_A||h_A)$ stored in the smart card. Unfortunately, $H(ID_A||RPW_A||h_A)$ is not utilized along Jeong et al.'s scheme, i.e. it is useless data. Therefore, GWN does not need to store the data when it generates a smart card. If that data is not in the smart card, $\mathcal{A}$ cannot correctly guess PW_A and RPW_A. Then the scheme provide security against privileged insider attacks.

C. Countermeasure of Impersonation Attacks

The adversary $\mathcal{A}$ can impersonate the user, because $\mathcal{A}$ can computes I_A. However, we mentioned, just before, that the scheme could provide security against privileged insider attacks and $\mathcal{A}$ cannot guess PW_A and RPW_A. Therefore, $\mathcal{A}$ cannot computes I_A as well.

D. Countermeasure of Session Key Disclosure Attacks

The session key is easily disclosed, because the random numbers k_A and k_S are encrypted with same key z_S. To prevent this attack, the random number should be encrypted with other data [19], [20] or another method to conceal data [21]

VI. Conclusions

User authentication and key agreement are important security requirements for IoT environments. And several multi-factor authentication schemes have been proposed in recent years. However, these schemes are vulnerable to quantum computing attacks and the security threats should be resolved. Recently, Jeong et al.'s proposed a NTRU-based user authentication scheme in IoT environments. They insisted that their scheme provides various security properties, even security against the quantum computing attacks. Unfortunately, we found out that their scheme has some incorrectness in authentication phase and security weakness against the privileged insider adversary. We presented the fixes for the incorrectness and the countermeasure for the security weakness. The scheme with the countermeasures provides a proper user authentication and security properties against various attacks.

For further works, we are designing completely a security-enhanced NTRU-based user authentication scheme in IoT environments.

Acknowledgment

This research was sponsored by 'QKD based next generation KREONET project' of KISTI.

References

[1] S. Challa, M. Wazid, A. K. Das, N. Kumar, A. G. Reddy, E. J. Yoon, and K. Y. Yoo, *"Secure signature-based authenticated key establishment scheme for future IoT applications"*, IEEE Access, 5, 3028-3043, 2017.

[2] M. Turkanovic, B. Brumen, and M. Hölbl, A novel user authentication and key agreement scheme for heterogeneous ad hoc wireless sensor networks, based on the Internet of Things notion, Ad Hoc Networks 20, pp. 96-112, 2014.

[3] X. Yao, X. Han, X. Du, and X. Zhou, A lightweight multicast authentication mechanism for small scale IoT applications, IEEE Sensors Jour., 13(10), pp. 3693-3701, Oct., 2013.

[4] B. Ndibanje, H. J. Lee, and S. G. Lee, Security analysis and improvements of authentication and access control in the internet of Things, Sensors, 14(8), pp. 14786-14805, 2014.

[5] P. W. Shor, *"Algorithms for Quantum Computation : Discrete Logarithms and Factoring"*, Proceedings of 35th Annual Symposium on Foundations of Computer Science and IEEE Computer Society, 124-134, 1994.

[6] L. K. Grover, *"A Fast Quantum Mechanical Algorithm for Database Search"*, STOC 96' Proceedings of the twenty-eighth annual ACM symposium on Theory of Computing, 212-219, 1996.

[7] ETSI, Quantum Safe Cryptography and Security, ISBN No. 979-10-92620-09-0, 2015.

[8] NIST, Report on Post-Quantum Cryptography, IR 8105, 2016.

[9] J. Hoffstein, J. Pipher, and J. H. Silverman, *"NTRU: A Ring-Based Public Key Cryptosytem"*, Algorithmic Number Theory, ANTS 1997, Lecture Notes in Computer Science (LNCS), Vol. 1423, 278-288, 1998.

[10] IEEE, *IEEE P1363.1 Draft 10: Draft Standard for Public Key Cryptographic Techniques Based on Hard Problems over Lattices, Internatiional Association for Cryptologic Research Eprint archive, 2008.*

[11] S. Roy, S. Chatterjee, and G. Mahapatra, *"An efficient biometric based remote user authentication scheme for secure internet of things environment",* Journal of Intelligent & Fuzzy Systems, 34(3), 1403-1410, 2018.

[12] K. S. Park, S. K. Noh, H. J. Lee, A. K. Das, M. H. Kim, and Y. H. Park, *"LAKS-NVT: Provably Secure and Lightweight Authentication and Key Agreement Scheme without Verification Table in Medical Internet of Things"*, IEEE Access, 8, 119387-119404, 2020.

[13] X. Li, J. Niu, S. Kumari, F. Wu, A. K. Sangaiah, and K.-K. R. Choo, *"A Three-Factor Anonymous Authentication Scheme for Wireless Sensor Networks in Internet of Things Environments*, Journal of Network and Computer Applications 103, 194-204, 2018.

[14] S. H. Jeong, K. S. Park, Y. H. Park, and Y. H. Park, *"An Efficient NTRU-Based Authentication Protocol in IoT Environment"*, Intelligent Computing, SAI 2018, Advances in Intelligent Systems and Computing, vol 857. Springer, 1262-1268, 2019.

[15] P. Kocher, J. Jaffe, and B. Jun, *"Differential power analysis"*, in Proc. 19th Annu. Int. Cryptol. Conf., Santa Barbara, CA, USA, Aug. 1999.

[16] S. J. Yu, J. Y. Lee, Y. H. Park, Y. H. Park, S. W. Lee, and B. H Chung, *"A Secure and Efficient Three-Factor Authentication Protocol in Global Mobility Networks"*, Applied Sciences 10, no. 10, 2020.

[17] S. J. Yu, K. S. Park, Y. H. Park, H. P. Kim, and Y. H Park, *"A Lightweight Three-Factor Authentication Protocol for Digital Rights Management System"*, Peer-to-Peer Networking and Applications, 2020.

[18] K. S. Park, Y. H. Park, A. K. Das, S. J. Yu, J. Y. Lee, and Y. H Park, *"A Dynamic Privacy-Preserving Key Management Protocol for V2G in Social Internet of Things"*, IEEE Access, 7, 2019.

[19] H. J. Lee, D. W. Kang, J. H. Ryu, D. H. Won, H. S. Kim, and Y. S. Lee, *"A Three-Factor Anonymous User Authentication Scheme for Internet of Things Environments"*, Journal of Information Security and Applications, 52, 2020.

[20] S. Ahmed, S. Kumari. M. A. Saleem, K. Agarwal, K. Mahmood, and M. H. Yang, *"Anonymous Key-Agreement Protocol for V2G Environment Within Social Internet of Vehicles"*, IEEE Access, 8, 2020.

[21] K. S. Park, S. K. Noh, H. J. Lee, A. K. Das, M. H. Kim, Y. H. Park, and M. Wazid, *LAKS-NVT: Provably Secure and Lightweight Authentication and Key Agreement Scheme Without Verification Table in Medical Internet of Things*, IEEE Access, 8, 2020.

Developing a Mining Robot for Mars Exploitation: NASA Robotics Mining Competition (RMC)

Tariq Tashtoush*[1], Agustin Velazquez[2], Andres Aranguren[3], Cristian Cavazos[4], David Reyes[5], Edgar Hernandez[6], Emily Bueno[7], Esteban Otero[8], Gerardo Zamudio[9], Hector Casarez[10], Jorge Rullan[11], Jose Rodriguez[12], Juan Carlos Villarreal[13], Michael Gutierrez[14], Patricio Rodriguez[15], Roberto Torres[16], Rosaura Martinez[17], and Sanjuana Partida[18]

School of Engineering
Texas A&M International University
Laredo, TX, 78041 USA

***Abstract*—This paper focuses on demonstrating the design and build stages, and effort done by Systems Engineering students team (DustyTRON NASA Robotics) to develop a mining robot that was used in the 2016 National Aeronautics & Space Administration (NASA) Robotics Mining Competition (RMC). The objective of the NASA RMC challenge is to encourage engineering students to design and build a robot that will excavate, collect, and deposit a simulated Martian regolith. Mining water/ice, and regolith is very essential task for space missions and resource utilization, they contain many elements such as metals, minerals, and other compounds. The Mining will allow extracting propellants from the regolith such as Oxygen and Hydrogen that can be used as an energy source for in-space transportation. In addition, the space mining system can be used in tasks that are important for human and robotics scientific investigations. The DustyTRON team consists of Systems Engineering students, who are divided into 1) hardware design, 2) electrical circuitry and 3) software development sub-teams. Each team works in harmony to overcome the challenges had previously experienced, such as heavy weight, circuitry layout design, autonomous and user control modes, and better software interface. They designed and built a remote controlled excavator robot, that can collect and deposit a minimum of ten (10) kilograms of regolith simulant within 15 minutes. The developed robot with its innovative mining mechanisms and control system and software will assist NASA in enhancing the current methodologies used for space/planet exploration and resources' mining especially the Moon and Mars. NASA's going-on project aims to send exploration robots that collect resources for analysis before sending astronauts. In 2016, only 56 United State (US) teams were invited to participate, and DustyTRON was one of three university teams from the state of Texas, the team placed the 16th in overall performance. This paper will address the full engineering life-cycle process including research, concept design and development, constructing the robot and system closeout by delivering the team's robot for the competition in Kennedy Space Center in Florida.**

Keywords*—NASA Robotics Mining competition; mining robot; ice regolith; autonomous; NASA space exploration; systems life-cycle; mechanical structure design; control system; systems engineering

I. Introduction

National Aeronautics and Space Administration (NASA) in the leader in space exploration the first robot landed on the moon. These robots were developed to be unmanned and will be the first to explore and mine resources from habitats that humans can not explore either due to high associated cost or highly hazardous planets ecosystem [1-11]. Robotics will allow humans to explore planets surfaces and resources while keeping high level of astronauts safety and lower costs to transport human to space. Additionally, robots will be capable of mining all the minerals and underground resources that will provide the needed energy (Oxygen and Hydrogen).

The NASA Robotic Mining Competition (RMC) encourages university students in the United States to be innovators and creative thinkers to design, build, and compete with robots that can traverse the simulated Martian chaotic terrain; then excavate the basaltic regolith simulant (called Black Point-1 or BP-1) and the ice simulant (gravel) , which are a representation of the necessary resource on Mars and return the excavated mass for a deposit into the collector bin to simulate an off-world mining mission.

On May 16-20, 2016, the seventh annual NASA Robotic Mining Competition was held at the Kennedy Space Center in Florida. This event brings together student teams from universities across the US to compete in a real of robotics, remote operation, and automation challenge related to NASA missions. Texas A&M International University (TAMIU) DustyTRON Robotic team, known as DustyTRON 2.0, worked to fulfill the competition goals, according to NASA's systems engineering guidelines and NASA RMC requirements [12-14]. This will mark the second entry into such competition and the team decided to build a new robot design from the ground up.

This paper is a detailed systematic engineering analysis and design process of DustyTRON 2.0 robot, where a fully functional mining robot will be constructed to fit certain specification including size dimension (1.5m * 0.75m * 0.75m), maximum weight (80 Kg), and mechanism (traverse, excavate and deposit). The paper will be centered on the team's design theory and Quality Function Deployment (QFD) analysis. Multiple designs are being evaluated according to preset criteria such as design to build, mobility, weight and budget. Then, one design will be chosen and analyzed in depth.

DustyTRON 2.0 had been divided into three sub-teams: 1) hardware design and construction, 2) electrical circuitry design, and 3) software development, in which the seventeen team members and there assignments is distributed as shown in Fig. 1.

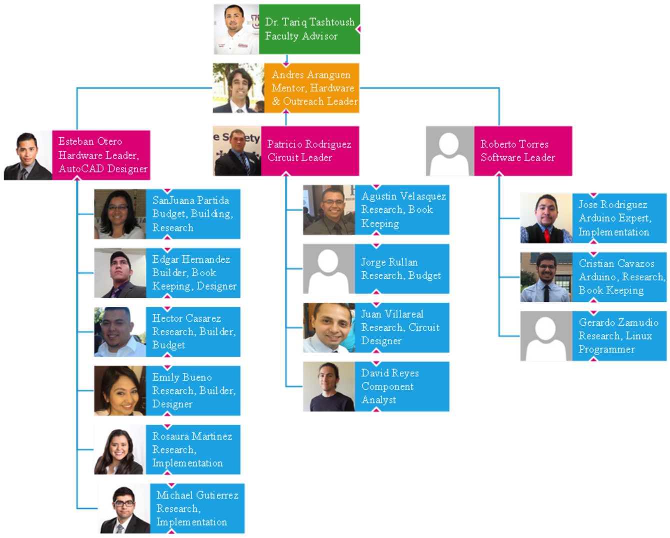


Fig. 1. DustyTRON 2.0 Team Management

The hardware design and construction team will be focusing on building a strong robot structure that can be moved easily keeping light weight while having an excavation, a regolith deposit and the moving mechanisms. The electrical Circuitry team is the bridge that links hardware and software together to make the robot functional. They will improve DustyTRON circuitry design to make sure that the components are easier to access. Cables will also be easily traceable in case of troubleshooting, thus the number of components required. While, the software development team intend on achieving autonomous functionality on DustyTRON 2.0 by utilizing the capability of a System-on-Chip (SoC) and microprocessor systems, where both systems will be communicating through serial interface. They will be using secure connection between the robot and the control station, while using Open-CV (Computer Vision) library for image and object detection to help achieving the autonomous mode. In addition, software development team will have to create an easy access and monitoring connection to the SoC system for rapid maintenance and troubleshooting immediately if errors occur during autonomous mode.

DustyTRON NASA robotics team participated in the NASA RMC 2015, with their mining robot DustyTRON 1.0, as shown in Fig. (2 and 3). There were several major issues with “DustyTRON 1.0” robot design that needed to be addressed and improved upon for the new design.

The first issue was the weight of the robot, DustyTRON 1.0 was borderline on the 80 kg, the weight limit requirement for the NASA RMC competition. That heavy weight altered the robot’s ability to move efficiently around the simulated terrain. For this reason, this year we only considered using only light materials when constructing the frame, the excavation mechanism, and the wheel system. DustyTRON 2.0 will be having lighter materials, a simpler frame, and lighter motors.

Second issue was DustyTRON 1.0’s steering system, it was not very efficient and resulted in limited maneuverability and agility which diminished the amount of runs the team was able to make during the allowed ten (10) minutes. The steering and wheel system included a tube shaped wheel that had a limited turning angle and because of the tube’s large surface area, and it had high friction due to large ground contact area. In addition, it had limited ability to drive over rocks and get out of any potential ditches.

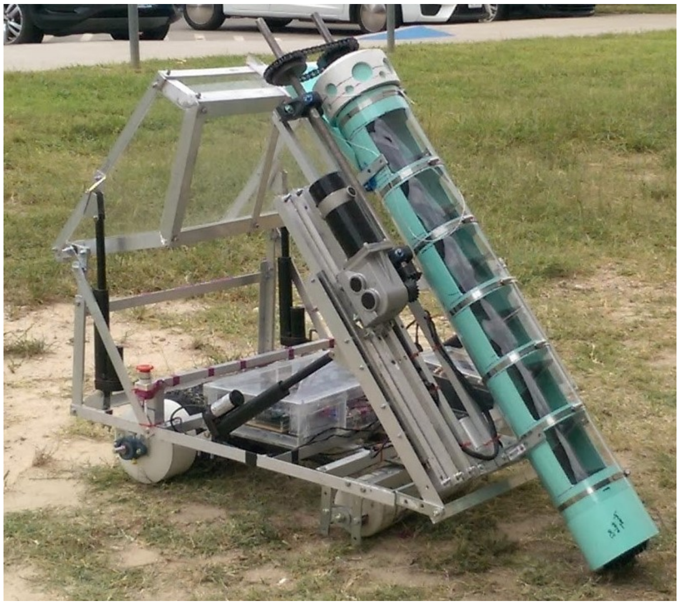


Fig. 2. DustyTRON 1.0 Robot

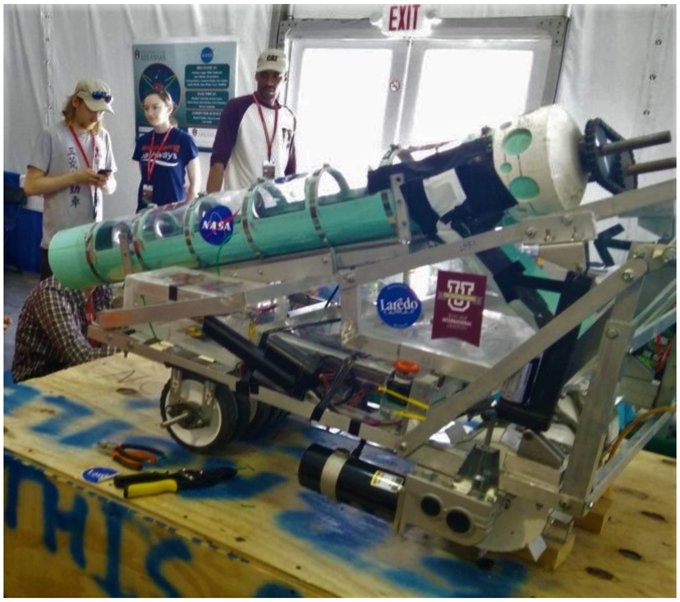


Fig. 3. DustyTRON 1.0 Robot Side-View

DustyTRON 2.0 will have an improved steering and wheels system by using four individual wheels were chosen as the new driving system with bigger tires that will be easier to manipulate and will traversed more efficiently through the simulated Martian terrain. These wheels would give DustyTRON 2.0 a clearance of six (6) inches from the ground to the lower digging mechanism (auger) tip and eight (8) inches for the frame while having the digging capability to adjust as desired using the adaptive suspension system. This space would be key in allowing the robot to drive over any potential small to medium sized rocks. Secondly, in conjunction with the software team, the circuitry team will be able to power and control each wheel-motor individually and independently, and adding better overall performance.

DustyTRON 1.0's upper frame was not secured enough and it would tremble excessively when the auger would excavate due to the fact of being top heavy, as shown in Fig. (4). This compromised the structural integrity of the whole robot. The DustyTRON 2.0 design will have a more stable frame that will allow it to remain stationary as the auger excavates and collects the regolith. Another improvement made to the newer design was to add a conveyor belt with scoops that would serve as both the collecting and dumping mechanism of the robot. This will reduce the weight of the robot and will simplify the control system.

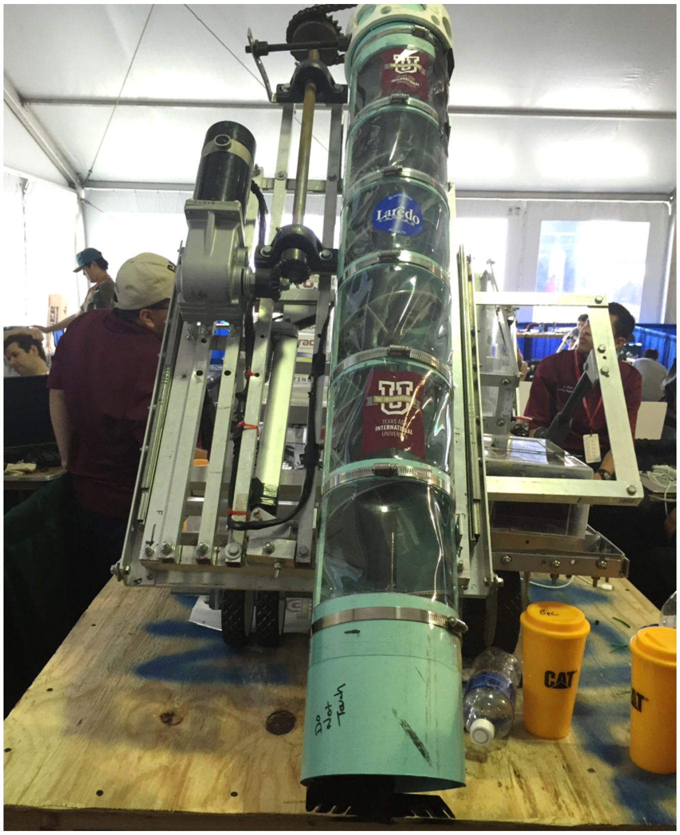


Fig. 4. DustyTRON 1.0 Robot Front-View

The electrical components in DustyTRON 1.0 robot were protected by a box made of Plexiglas, which was located under the excavation mechanism. This made it difficult to access, modify, or replace any of the components without repositioning the robot. Another drawback of the design was the way in which components were placed near areas of robot high activity, where many parts move directly over and regolith passing above it. DustyTRON 2.0 design of the box will be made more ergonomic. The electrical layout and connections are being placed in a more organized manner. The overall circuit design had been improved to include electrical box located at an elevated position and away from moving components, such as the auger and dumping mechanism. The placement at a higher point gives the DustyTRON 2.0 team easier access for maintenance, troubleshooting, or parts replacement.

DustyTRON 1.0 robot was controlled remotely using two Xbox 360 controllers that were used to control directional movement and the excavation mechanism separately. DustyTRON 2.0 will attempt the autonomous control with the option of manual control with two controllers. Another addition to DustyTRON 2.0 are optical sensors such as a Microsoft Xbox One Kinect and a rear local network (IP) servo-controlled camera. These will allow the robot to work autonomously which will require less human intervention and less bandwidth usage. The Kinect camera is installed in the front part of the robot to allow it to scan the environment. The rear servo-camera is installed to allow the monitoring and regulation of both collection and deposit mechanism. These cameras are new components that DustyTRON 1.0 did not possess, it only relied on manual control and simulated terrain cameras. In summary, these components will allow DustyTRON 2.0 to be independent, achieve autonomous, manual control and use less bandwidth and power.

The team felt there were opportunities to improve our robot's functionality and coding by cleaning up the Arduino code. This simplified the code and facilitate the command and communication processes. Another way that DustyTRON 2.0 is trying to fulfill the requirements of being autonomous mode is by using the Jetson TK1. This will aid in wireless communication and computer vision processing capability via the Graphics Processing Unit (GPU). Software Team worked together when coding the Arduino in order to ensure everyone would have their input in the coding and understand when shifting the workload between the Arduino Mega and the Jetson TK1.

DustyTRON 2.0 team cleaned up the coding of the manual controlling Visual-Basic code. The manual mode will be utilized in the event that the Xbox camera would not function properly for the image detection, object avoidance or in the event of any other connectivity issues. The software team is planning to establish an algorithm for the usage of the Xbox 360 wired controllers. The establishment and application of these algorithms will allow for full control of the robot's wheel motors with the controller. It will also provide direct commands to the robot's functions.

The paper is organized as follows: Section 2 is short history of NASA mission to explore the Moon and Mars and how it was done, Section 3 a summary of system requirements and the team's preliminary Designs, Section 4 describes concept operation, Sections 5 illustrates the system hierarchy, while Section 6 Risk management, trade-off analysis, verification of System Meeting Requirements and reliability, Section 7 summarises the results of the competition, and Section 8 concludes the paper and describe the team future plan.

II. Literature Review and NASA RMC History

This section aims to introduce and explain how the NASA Robotic Mining Competition (RMC) works. Recent discoveries by NASA missions to Mars have found large amounts of water in the form of water ice and hydrated minerals on Mars utilizing the space rovers such as "Curiosity" and orbiting satellites [12-14]. These sources of water resulted from clay and clay-like minerals that formed millions years ago on the surface or underground of Mars. Capturing this water is key to allow humans to “live off the land” by utilizing the available resources. The water can be used for human use, hygiene, rocket propellant, growing plants, radiation shielding, and can be used in various other processes. in order to gain access to that water, minerals that contain the water must be mined out and the surface soil “regolith” be removed to expose the water resources.

The NASA Robotic Mining Competition is for university-level undergraduate students to challenge them to design and build a mining robot that can traverse simulated Martian terrain. The mining robot must excavate the regolith simulant and/or the ice simulant and return and deposit the excavated mass to a collection bin from a space station. The challenge contains few complexities such as the complexities the abrasive characteristics of the regolith, the limited robot weight/size and the required tele or autonomous operation of the the robot from a remote Mission Control Center. Additionally, participating teams must consider a number of design and operation factors such as dust tolerance and dust projection, mass, communications/energy/power consummation and autonomy.

NASA benefits from the competition by encouraging student teams to develop innovative robotic excavation concepts, which may lead to a creative ideas that can be used in an actual excavation device for NASA future missions. Advances in Martian mining have the potential to significantly contribute to human spaceflight and NASA space exploration operations. Details of this competition can be found online at https://www.nasa.gov/nasarmc.

The NASA RMC started in its original format in 2010 as NASA Lunabotics Competition [13]. In 2011, it was open to undergraduate and graduate student teams enrolled in colleges or universities worldwide. But in 2014, due to NASA budgetary constraints, the competition was limited to teams from United States colleges or universities. In 2020, NASA transited to a Lunar focused competition, and Table I represent the competition year, name and the allowed countries to participate.

TABLE I. NASA Robotics Mining Competition History

Competition Year and Name	Competition Participants
(2010) Lunabotics	USA
(2011) Lunabotics	USA, Bangladesh, Canada, Colombia, India, Spain
(2012) Lunabotics	USA, Bangladesh, Canada, Colombia, India, Mexico, Romania, South Korea
(2013) Lunabotics	USA, Australia, Bangladesh, Canada, Colombia, India, Mexico, Poland
(2014-2019) RMC	USA
(2020-present) RMC: Lunabotics	USA

Many teams allover the United state presented in the robotics design they came up with and met the NASA RMC requirements [15-22]. In this paper DustyTRON team is presenting their robot, the design process and their achievements in the 2016 NASA RMC. This paper will provide an insight for other researchers and teams to follow the system engineering concepts and process to develop new technologies and designs that can improve the human race to explore and populate space. This paper is a successful example of utilizing and implementing systems engineering concepts to real-life problems and industries not traditionally known for the use of systems engineering.

III. System Requirements and Preliminary Designs

The purpose of the project is to create an inexpensive improved rover system that can perform multiple functions, such as image capturing, rock mining, and data collection while exploring the space. Many researchers and engineering teams [23-34] worked on developing new technologies that can help to explore beyond our earth mainly Mars and the moon.

A. System Requirements

To begin the design process, the team gathered the requirements that will benchmark their design. These requirements were derived from the NASA RMC competition rules and regulations, and it meant to ensure meeting the competition needs and goals. The team broke the system down into functional subsystems and identified how they would interact. Then, the team generated concepts for each subsystem, scoring them against the requirements to determine the final design. The key requirements are listed in Table II.

TABLE II. System requirements extracted from [23-25]

Requirement Type	Action	Specifications
Performance Requirements	Regolith: Excavate	Have an excavation mechanism that will be able to excavate an adequate depth that will reach the ice simulant
	Regolith: Collect	The robot must be equipped with a form of storage that it will use to collect the regolith excavated
	Regolith: Deposit	The robot must be capable of depositing the regolith collected onto a bin located at the back of the simulated terrain
Design Requirements	Dimensions	The robot must not exceed the measurements of 1.5m in length, and 0.75m in both height and width
	Weight	The robot must weigh less than 80 Kg in order to compete

B. DustyTRON 2.0 Preliminary Designs

1) Design #1: The first robot design shown in Fig. 5, illustrate a robot that will excavate the simulated Martian regolith using two augers and store the regolith into a Plexiglas box, as shown in the AutoCAD design. This design will feature a frame that will be built using aluminum flat bars and angles and with PVC pipes to cover the augers. The chassis design is 1 m in length and 0.5 m in width.

The chassis will contain two Plexiglas boxes, one for the micro-controller, batteries, motors drivers and other electronics compartment, while the second box is for regolith collection behind the auger system. The four actuators, suspension, and wheels will be mounted to the chassis. Front wheels will be four inches in diameter and an inch wide, while rear wheels will be ten inches in diameter and four inches wide.

The auger system consist of two augers that will have a length of 0.513 m, and each auger will be housed in a 0.152 m inner diameter PVC pipe. The augers will be attached to the chassis, which acts as the base of the robot, and will be manipulated by the moving of the four cornered actuators/suspensions. The system will be powered with one 24V 280W motor.

The team decided to make the rear wheels larger and wider to help the robot maneuver more efficiently on the simulated Martian terrain. Two augers were added in order to increase the amount of regolith that could be excavated during the allowed ten minutes. The team decided to utilize Plexiglas again for the electrical box since it is durable and lightweight material, and has been proven to be able to protect the electrical components quite efficiently. The location of the box was also a major concern, as it needed to be easily accessible.

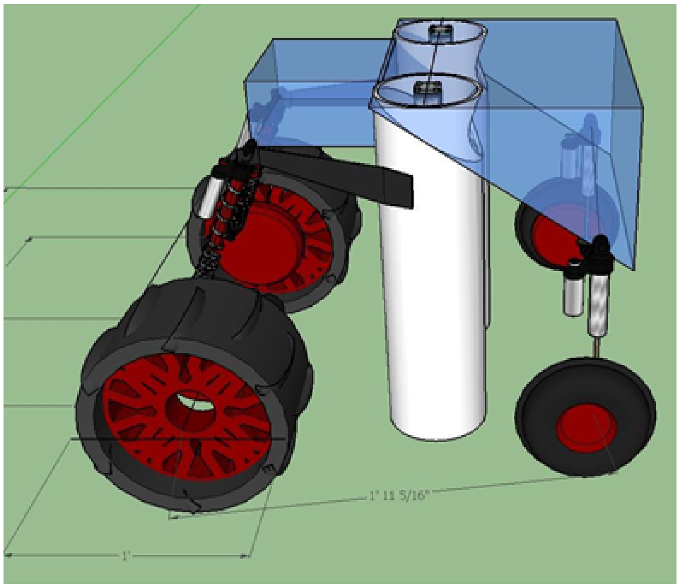


Fig. 5. DustyTRON 2 Mechanical Structure AutoCAD Design 1

2) Design #2: The structure for design 2 is very similar to that of Design 1 and shown in Fig. 6, except for a few changes that team members felt that they were necessary. Firstly, the team decided to only use one auger. This was in part due to the fact that one auger would still excavate an adequate amount of regolith, and the addition of another auger would only increase the power consumption. Another component that differs from Design 1 is the wheel system, which consists of six wheels that will be eight inches in diameter and six inches wide. The team also decided that the center wheels will be the only ones with a motor, this wheel setup requires two 12V motors.

3) Design #3: The third design structure is similar to the prior designs, except for a few differences. The number of wheels was decreased to four, and the number of actuators was decreased to only two, unlike the last two designs which included four actuators. The size of the wheels will be larger in order to provide better traction and mobility to the robot. Also, the collection bin will be reinforced with a metal frame in order to attach both a camera and emergency stop buttons. The illustration for this design is show in Fig. 7.

4) Design #4: For design 4 illustrated in Fig. 8, the upper structure underwent significant modifications, mainly the use of a new lifting mechanism. The team decided to implement a scissors-lift mechanism to the robot, instead of actuators as a mean to lift the upper structure. This design will allow the robot to be able to excavate deeper, but it will also have to support all of the robot components excluding the wheels. The number of actuators for this design remained as two. All other aspects of the design were the same as previous designs including the wheels, collection bin, camera, and the electrical components box.

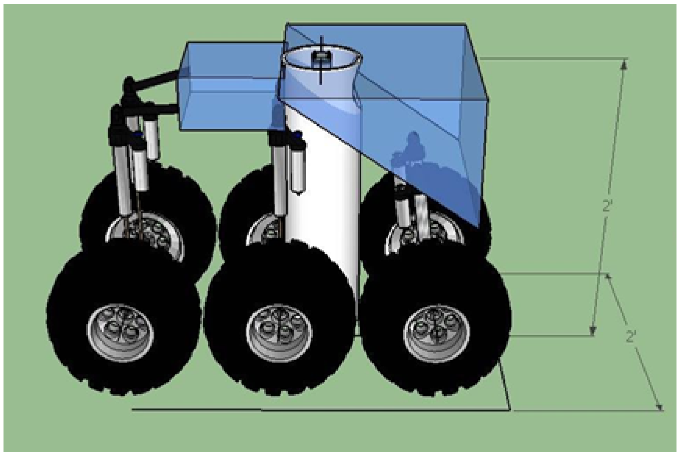


Fig. 6. DustyTRON 2 Mechanical Structure AutoCAD Design 2

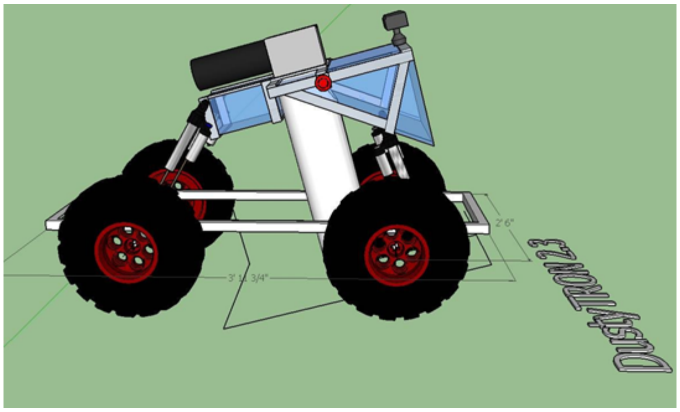


Fig. 7. DustyTRON 2 Mechanical Structure AutoCAD Design 3

5) Design #5: Final Design: This design underwent various major changes, the first and most important one is the addition of a slider to the excavation mechanism. This means the auger, and the powering motor will be mounted on a sliding mechanism that will allow for deeper reach. Another major change was the to change the frame materials from aluminum bars to 80/20 T-slotted extrusion bars. This material was the only material used to build both the frame and the sliding mechanism that would hold the auger and its motor. Another major addition was a conveyor belt equipped with scoops that will act as both a collecting box and dumping mechanism. The belt will collect the dirt and hold it until an adequate amount is collected, then it will deliver the regolith to the collecting bin. This ultimately became the final design that was built because of many reasons, the most important one being the weight. The t-slotted bars are both strong and extremely lightweight, making the frame both stable and light. Also, the

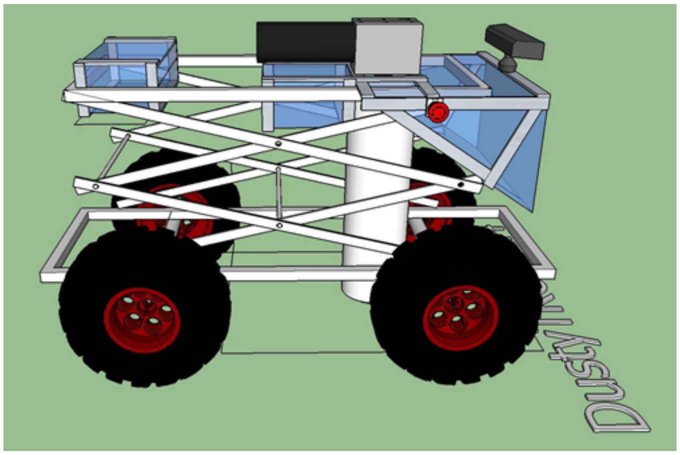


Fig. 8. DustyTRON 2 Mechanical Structure AutoCAD Design 4

sliding mechanism proved to be a better option to increase the depth of the excavation without compromising the integrity of the frame. This design is shown in Fig. 9.

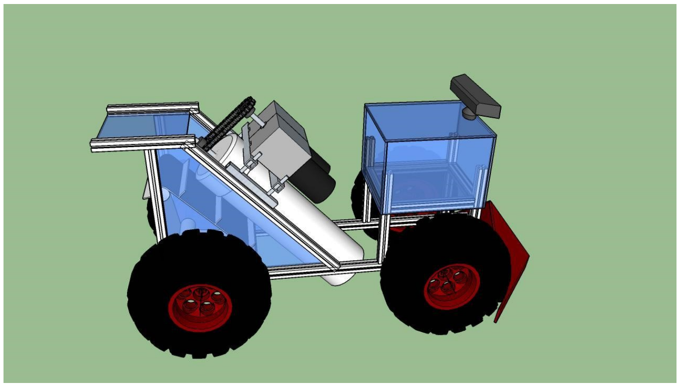


Fig. 9. DustyTRON 2 Mechanical Structure AutoCAD Design 5

Per DustyTRON 2.0 design, the system required four independently moving wheels and one 6 inches auger to excavate. These requirements laid out the fundamental ideas of circuit design. It was immediately recognized that four 12V motors were going to be required for the wheels as well as one 24V motor for the auger, these requirements helped in defining the power distribution setup. Using 12V motor and motor drivers in the testing stage, showed the power was not enough to rotate the auger at the desired speeds. Therefore, circuit team decided to use the same batteries connection as DustyTRON 1.0 but with the option of having both 12V and 24V outputs. Working with four batteries, two pairs would be connected in series and the pairs would then be connected in a parallel configuration, which will provide each component with the appropriate voltage, either 24V/14A or 12V/7A.

DustyTRON 2.0 design required different software configuration and code that will make a simpler operation. To achieve that, Arduino codes have been reformed (easier to read and user-friendly) to make executable commands easier and more precise. This is a significant improvement over the previous DustyTRON Arduino code, which was too complex and only readable to DustyTRON 1.0's programmers. Software team worked collectively to code Arduino and each of its individual parts to be readable to every team member. In addition, Arduino open source libraries will be used for each of the individual components of the robot, such as Victor motor drivers, Jaguar motor drivers, and Axis 206 local network (IP) Camera are all using the Servo library from Arduino. While the Pololu motor driver will be using last year's library to operate the auger motor and the linear actuators.

IV. Concept Operations

DustyTRON 2.0 must meet the required performance of the RMC and outperform last year's robot. Its frame will be substantially lighter than the allotted 80 Kg. The frame is constructed using 80/20 T-slotted extrusion bars, which is lightweight and will be used for different aspects of the robot. To excavate the required BP-1 plus the gravel (icy regolith simulant); DustyTRON 2.0 is using a shorter auger, which is powered with two 24V CIM motors with gear box of 47:1 ration.

DustyTRON 2.0 will feature a double spiral auger to increase productivity in the same amount of time. It will also be operated autonomously through multiple cameras, microprocessor and graphical processing unit, and if that fails, it will be manipulated over WiFi by two controllers. The robot will also incorporate a conveyor belt that will store then transport the collected BP-1 into the deposit bin. This will help in simplifying the collecting process by reducing the auger movement, and eliminate the storage box. Also, four 30 cm wheels will help to overcome any obstacle in the simulated Martian terrain while a front guard will also protect all the component under the robot.

V. System Hierarchy

To effectively visualize the main components of DustyTRON 2.0 robotics and their tasks and relation to each other, a system hierarchy was mapped out for each sub-team. The hardware components' hierarchy is showing in Fig. 10, it represent the mechanical components that dictate the robot structure integrity and performance. Robot structure consist of the wheels (for linear motion and steering), linear actuators (for individual excavation, collection and depositing components movements and suspension/steering system) and Motors.

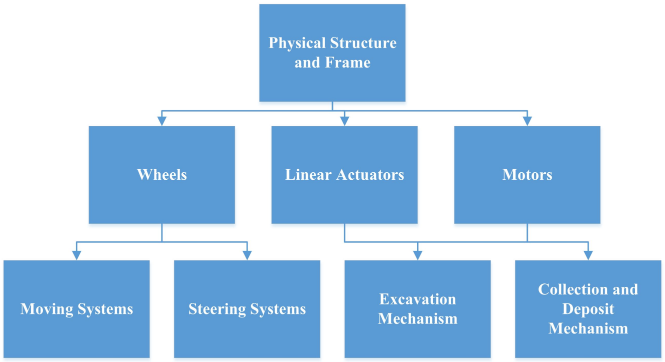


Fig. 10. Hardware System Hierarchy for DustyTRON 2.0

For the electrical circuit and software sub-teams, many parts are common between them. For manual control, a Microsoft Xbox 360 controller sends command inputs through the controller computer, which is wireless connected to the NVIDIA Jetson TK1 that's embedded in DustyTRON 2.0. For autonomous mode, a Microsoft Xbox One Kinect feeds image and video data to the TK1 for objects detection purposes, while a rear servo camera feeds image and video data to the TK1 in regards to regolith deposit. Whether it be manual or autonomous mode, TK1 will use the provided input to command to the Arduino Mega, which will be directly controlling the motor drivers. These motor drivers will power all motors to achieve the desired movements as shown in Fig. 11.

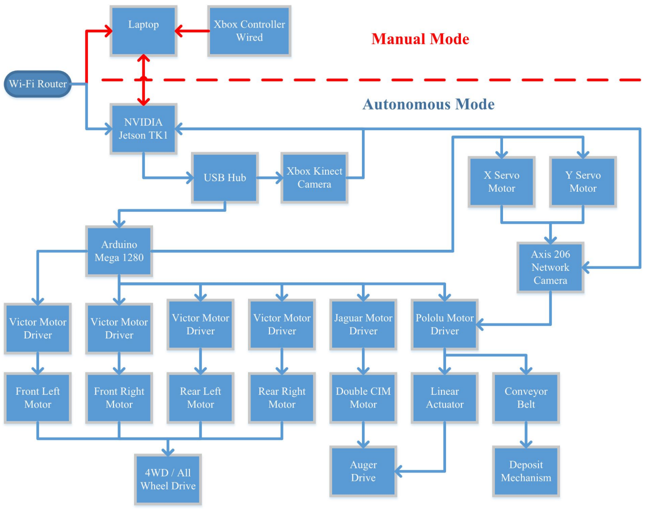


Fig. 11. Circuit System Hierarchy for DustyTRON 2.0

Software team decided to use SSH (Secure shell), which is an encrypted network protocol, and prevent unauthorized access to the Jetson TK1. In addition to using Virtual Network Computing (VNC) communication, bandwidth can also be reduced by compressing the video in the TK1 before being sent to the client computer. The TK1 will use a USB port for both serial communication and powering the Arduino, which will eliminate the need for an extra battery for the Arduino. When the Kinect creates a 3D environment for the TK1, the Kinect will allow the code to choose the shortest path to the target by using the "A Star Algorithm" and drive around any obstacles processes, Fig. 12 is showing the software system hierarchy.

Autonomous C++ code will be built in order to send commands to Arduino and control all motors for a successful move through the Martian simulated terrain; Fig. 13 represent the control setup for the Arduino microprocessor.

There were three types of interfaces that occurred; mechanical interface, electrical voltage/current interface, and data/digital interface. The first one is the hardware components interacting with each other such as gears and chains. The second is hardware components communicating with electrical component which will be dominate with electrical voltage/current signal. The third interface is digital data which occurs between the GPU/processor, sensor, and cameras. Some of the interfaces that will be utilizing for the software development of the robot include Ubuntu 14.04, Arduino Software IDE and Visual-Basic. The Ubuntu interface was used in order to install, update, and operate the Jetson TK1. The Arduino Software interface was used in order to code the Arduino Mega. Visual-Basic was used in order to program the control configuration for both of the wired Xbox 360 controllers. These interface and interactions between all DustyTRON 2.0 robot's systems is displayed in Fig. 14.

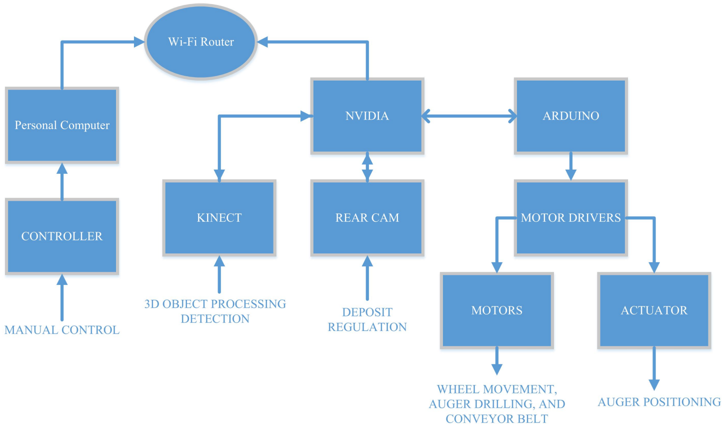


Fig. 12. Software System Hierarchy for DustyTRON 2.0

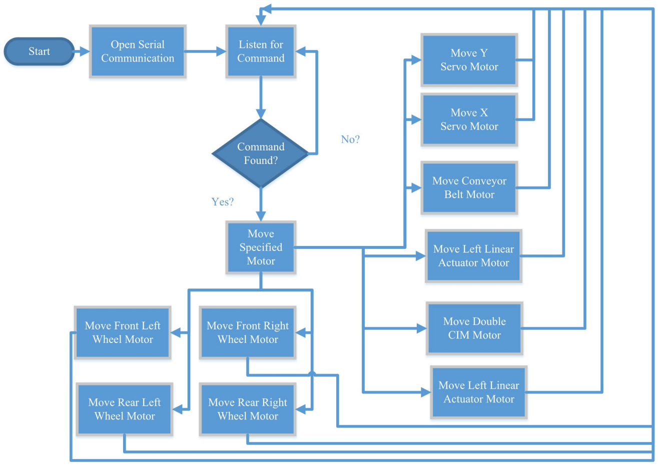

Start
Open Serial Communication
Listen for Command
Command Found?
No?
Yes?
Move Specified Motor
Move Y Servo Motor
Move X Servo Motor
Move Conveyor Belt Motor
Move Left Linear Actuator Motor
Move Double CIM Motor
Move Left Linear Actuator Motor
Move Front Left Wheel Motor
Move Front Right Wheel Motor
Move Rear Left Wheel Motor
Move Rear Right Wheel Motor

Fig. 13. Arduino Control System Hierarchy for DustyTRON 2.0

VI. Risk Management, Trade-off Assessments and Requirements Verification

To be prepared in case anything does not go according to plan, the team came up with various ways the robot would fail. Each failure was ranked according to the consequences on the overall performance of the robot, also, each failure is given a likelihood of happening.

A. Hardware Team

1) Risk Management: Fig. 15 was made to illustrate both the importance and probability of each failure (risk) for robot hardware design during or before the competition:

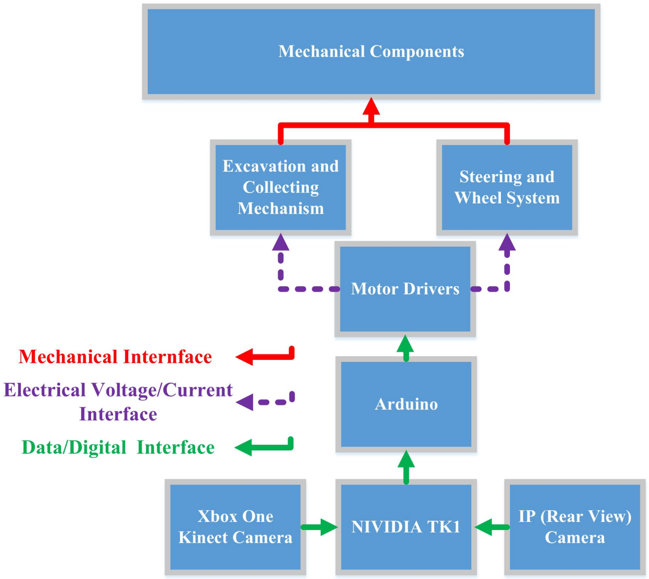
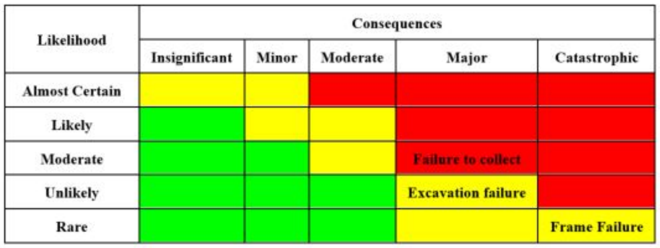


Fig. 14. Interaction Hierarchy between DustyTRON 2.0 Systems

- Failure of the excavation system: If the excavation system is not able to function, or fails to excavate enough regolith to satisfy the requirements.
- Frame integrity failure: If the frame is not able to support the weight of the auger, electrical component box, or the weight of the collected regolith.
- Failure to collect regolith: This might happen if the conveyor belt is not able to hold the regolith securely, or if the auger fails to deliver the regolith to the conveyor belt.

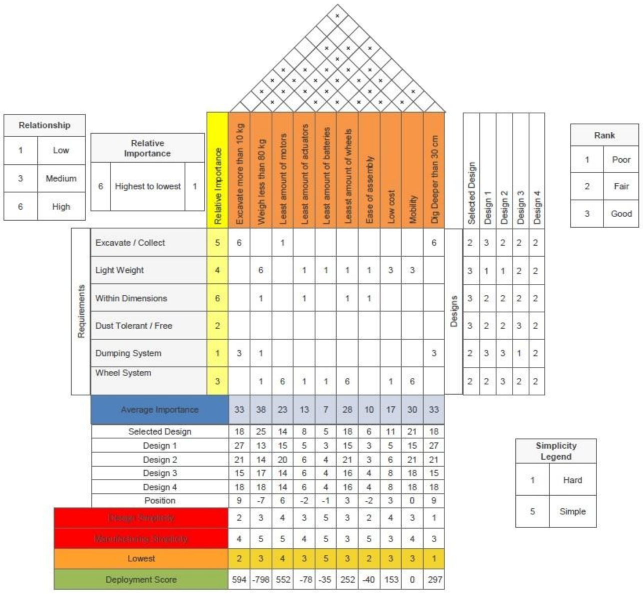
By categorizing the type of risks, and the possibility of them happening, the team will be able to better adapt to the failures and have be prepared to fix any issues that might arise.



Likelihood	Consequences				
	Insignificant	Minor	Moderate	Major	Catastrophic
Almost Certain	🟨	🟨	🟥	🟥	🟥
Likely	🟩	🟨	🟨	🟥	🟥
Moderate	🟩	🟩	🟨	🟥 Failure to collect	🟥
Unlikely	🟩	🟩	🟩	🟨 Excavation failure	🟥
Rare	🟩	🟩	🟩	🟨	🟨 Frame Failure

Fig. 15. DustyTRON 2.0 Hardware Risk Management

2) Trade-off Assessment: Quality Functional Deployment (QFD) is a focused methodology that takes into account the voice of the customer and develop a response to those needs and expectations. The customer is NASA and they provided the team with their voice as rules and regulations, the team developed few capabilities that can be controlled to achieve the needs. Fig. 16 shows the QFD developed including a list of needs was generated based on the criterion that was set by NASA's Robotic Mining Competition. Then a list of specifications based on these needs and their importance was created and our team goals were included and prioritized according to their importance based on the performance improvement required. This method allowed the team to weight their designs and found which one should be selected which was our final design.



Relationship	
1	Low
3	Medium
6	High

Relative Importance		
6	Highest to lowest	1

Requirements	Relative Importance	Excavate more than 10 kg	Weigh less than 80 kg	Least amount of motors	Least amount of actuators	Least amount of batteries	Leasst amount of wheels	Ease of assembly	Low cost	Mobility	Dig Deeper than 30 cm	Selected Design	Design 1	Design 2	Design 3	Design 4
Excavate / Collect	5	6		1							6	2	3	2	2	2
Light Weight	4		6		1	1	1	1	3	3		3	1	1	2	2
Within Dimensions	6		1		1		1	1				3	2	2	2	2
Dust Tolerant / Free	2											3	2	2	3	2
Dumping System	1	3	1								3	2	3	3	1	2
Wheel System	3		1	6	1	1	6		1	6		2	2	3	2	2
Average Importance		33	38	23	13	7	28	10	17	30	33					
Selected Design		18	25	14	8	5	18	6	11	21	18					
Design 1		27	13	15	5	3	15	3	5	15	27					
Design 2		21	14	20	6	4	21	3	6	21	21					
Design 3		15	17	14	6	4	16	4	8	18	15					
Design 4		18	18	14	6	4	16	4	8	18	18					
Position		9	-7	6	-2	-1	3	-2	3	0	9					
Design Simplicity		2	3	4	3	5	3	2	4	3	1					
Manufacturing Simplicity		4	5	5	4	5	3	5	3	4	3					
Lowest		2	3	4	3	5	3	2	3	3	1					
Deployment Score		594	-798	552	-78	-35	252	-40	153	0	297					

Rank	
1	Poor
2	Fair
3	Good

Simplicity Legend	
1	Hard
5	Simple

Fig. 16. DustyTRON 2.0 Quality Function Deployment (QFD)

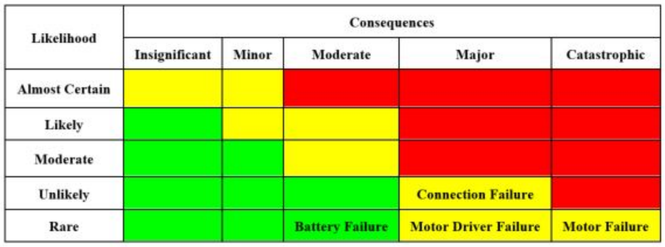
Comparing the critical design to preliminary designs, robot dimensions/weight was ranked the highest priority task. Therefore, an option of balancing the NASA RMC constraints and the team goals were selected. Some of the trade-off for the hardware team include not using an adaptive suspension system to change the robot height and stabilize the robot frame on a rough terrain, because of the added weight. Also, switching from a bigger collection box to conveyor system to simplify the electrical-mechanical components.

B. Electrical Circuit Team

1) Risk Management: Circuits can be very delicate but can be made to withstand the system requirements. As with every component of the robot, circuit design was also taken into consideration for any possible failures. Fig. 17 categorizes the possible circuit risks for the robot:

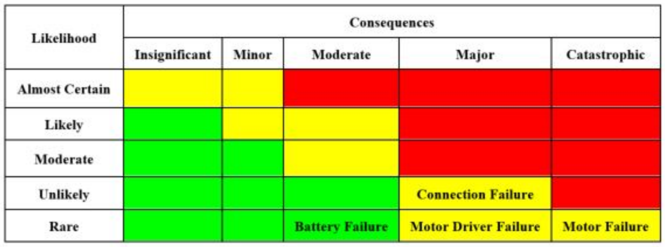
- Battery Failure: If one of the batteries fails to provide the required voltage/current.
- Connection Failure: If connections get lose or overheat.
- Motor Driver Failure: If motor driver is not powering its assigned motor any longer.
- Motor Failure: If the motor malfunctions and will not be able to drive the necessary component (wheel or auger).

Because these problems may show during the building process, circuit team prepared for every scenario during the



Likelihood	Consequences				
	Insignificant	Minor	Moderate	Major	Catastrophic
Almost Certain	🟨	🟨	🟥	🟥	🟥
Likely	🟩	🟨	🟨	🟥	🟥
Moderate	🟩	🟩	🟨	🟥	🟥
Unlikely	🟩	🟩	🟩	🟨 Connection Failure	🟥
Rare	🟩	🟩	🟩 Battery Failure	🟨 Motor Driver Failure	🟨 Motor Failure

Fig. 17. DustyTRON 2.0 Electrical Circuit Risk Management



Likelihood	Consequences				
	Insignificant	Minor	Moderate	Major	Catastrophic
Almost Certain	🟨	🟨	🟥	🟥	🟥
Likely	🟩	🟨	🟨	🟥	🟥
Moderate	🟩	🟩	🟨	🟥	🟥
Unlikely	🟩	🟩	🟩	🟨 Connection Failure	🟥
Rare	🟩	🟩	🟩 Battery Failure	🟨 Motor Driver Failure	🟨 Motor Failure

Fig. 18. DustyTRON 2.0 Software Risk Management

testing phase to prevent any future possible problem. Selecting compatible components, and verifying its operation which includes correct wiring under the correct conditions, the probability of risk is greatly reduced. However, in case of unfortunate situation of a component breaking down, team designed the circuit in a simple way that will allow easier components replacement.

2) Trade-off Assessment: The first trade-off was the power source, originally, the robot was running a high discharge 22.2V (5000 mAh) 6-cell battery which was super light (1.5 lb.) but it was extremely powerful and was able to burn several motor drivers. Therefore, team decided to go with multiple lower power batteries even if they are heavier (16 lb.) and required more connections. The second major trade-off was the motor drivers, the first selected type was an easy to plug and play devices but they have a limited performance. The team opt to use different brand which required soldering and configuration the motor driver to the required setting because of the superior performance.

C. Software Team

1) Risk Management: Software failure can be due to either connection of related-hardware failure. Software team failures might be:

- Failure of NVIDIA TK1 power regulator: If power regulator of the TK1 is not able to function properly, the TK1 will be damaged and not work at all.
- Failure of VNC connection: If the VNC connection were not working properly, no rapid maintenance will be provided to the TK1 thus a risk for potential inoperability.
- Failure of programming Open-CV: If Open-CV library were not configured or programmed correctly, the robot autonomous mode will be disabled.
- Failure to send command to Arduino: If we were to run out of characters to program each movement of the robot, we would not be able to move the robot's individual components.
- Failure to have available PWM pins in Arduino: If we run out of PWM pins, we would have to change the code to fit analog and digital pins instead.

These failures had been illustrated in Fig. 18.

2) Trade-off Assessment: Software team had a major trade-off issue by using Jetson TK1 with lower processing power over Jetson TX1 because TX1 was back-ordered. For manual control, they opt to use Xbox 360 wired controllers to avoid any lag issues and wire management will be required. In addition, two Xbox 360 controllers will be used, where one person can excavate while the other person operates the robot. This will limit the controller buttons mapping errors and reduce the human error. The button layout of the Xbox 360 controller is arguably the best ergonomic and user friendly in the market. Arduino Mega 1280 was used over Arduino Mega 2560, because Arduino Mega 1280 is available from last year which will reduce the budget. This trade-off resulted in reducing the available memory but the robot gained an extra 20 mA DC current per I/O pin. Arduino Mega will be used for Pulse-Width-Modulation (PWM) instead of Jetson TK1 because Arduino is more stable. Xbox Kinect Camera was selected over other cameras because of its 3D processing, sensors, and image detection capabilities, which will benefit in achieving the autonomous mode. Axis 206 Camera was used over other cameras because it comes with two servo motors to control x- and y-axis.

D. Verification of System Meeting Requirements

To create DustyTRON 2.0, the team felt the robot had to meet certain specific requirements which were both team-oriented and NASA RMC derived. The following requirements were taken into consideration.

1) Functional Requirements: Robot will be able to excavate the simulated Martian terrain.

- Robot shall be able to operate by tele-robotic operations or autonomously to reach its destination.
- Collected material shall be deposited and stored into the conveyor belt until the deposition.
- The robot shall travel the arena to the collector bin and if needed return to the mining area to continue excavating.
- When collecting and depositing the obtained BP-1, robot shall be able to evade the obstacle presented in the arena.
- Easier and understandable code shall be established for the usage of controllers.

2) Performance Requirements:

- Robot shall be able to collect 10 kg of BP-1 to deposit in the collector bin and shall be able to operate for 10 continuous minutes.
- The BP-1 excavated and collected shall be obtained from opposite end of the arena from where the robot is placed and robot shall begin excavating once it reaches a mining line.
- Robot shall incorporate a protective mechanism to shield the electronics parts and avoid picking up excess amount of dust.
- The bandwidth that the robot consumes during communication functions shall be reduced.

3) Physical Requirements:

- Mining robot shall weigh a maximum of 80 kg including any subsystems and cameras.
- Robot shall be self-powered with on-board power and the energy used shall be recorded.
- Initial dimensions of the robot shall be within 1.5m, 0.75m, and 0.75m in length, width, and height, respectively.

4) Safety Requirements:

- An emergency red button will be installed on the robot to terminate its actions. The button shall be 40 mm in diameter or greater and should be easily accessible.
- The electrical wiring must be correctly installed with the appropriate connections to avoid any accidents.
- An easy connection setup between the robot and the control station, while keeping a high level of security, shall be incorporated.

To ensure that the robot's system is meeting all necessary requirements, the team must follow NASA RMC's regulations and guidelines. Verification started by testing and inspecting each component for defects. Commercial off-the-shelf (COTS) parts were tested before used on the robot to verify their performance and specifications. Some of the actions that DustyTRON 2.0 team followed are:

- The extruded T-slot bars' integrity was inspected for any damage or defects before constructing the frame. Fitting and testing bar ability to support the weight of the excavation mechanism and its vibration.
- Twin spiral auger was tested in the simulated field, to make sure it would reach an adequate depth in order to reach the ice region while being able to excavate an adequate amount of regolith.
- Linear actuators were tested to be able to lift and push the weight of the auger mechanism and frame.
- Excavating motor was tested to make sure it can produce enough torque to satisfy the excavating needs, break through the surface of the simulated terrain, and handle the ice-simulated material.
- Conveyor belt was verified to be strong and sturdy to not come apart once it starts moving while carrying the regolith or ice simulant. Also, the smooth and stable movement of the conveyor rollers and motor was tested.
- Wheels and their motors were tested on sand in order to verify the ability to maneuver effectively through such terrain carrying the robot and collected regolith weight.
- Batteries were tested to make sure that they have the required power to operate the robot with all of its components at maximum capacities for more than 10 minutes.
- Emergency-stop buttons were tested to verify if they could cut off the batteries' power from the whole robot when needed.
- Power consumption analyzer were tested to confirm its calibration status.
- 4-Wheel system was programmed and tested using the selected motor drivers and the Xbox controllers.
- Axis 206 Network Camera were tested and verified that it can reach the 180 degrees rotation limit while broadcasting the live feed to NVIDIA TK1.
- Xbox Kinect V2 camera functionally was validated by processing live images and to detect object depth perception.
- Stable VNC connection over the WiFi was established and tested to allow a complete control and monitoring of the TK1 through a laptop to provide maintenance when needed.

E. Reliability

To ensure the maximum reliability of the robot, DustyTRON 2.0 team has taken precautions in solving the potential issues from last year's robot, which will help to elongate the robot's Mean Time Between Failures (MTBF). Hardware team focused on building a frame that is reliable enough to support the excavation mechanism, along with the collected regolith weight, while still being able to move around efficiently by selecting better wheels to maneuver in the uneven simulated terrain during the competition.

For the circuit team, reliability will be increase by reducing the amount of cables and connection used, because less components means less probability of having a breakdown. The second task is securing and protecting the cabling from harm or damaging elements. Cable management will be taken into account by using power distributor board with attached voltage regulator, which will eliminate the need of any unnecessary connection with multiple voltages. The robot's power source will be four batteries which are connected in serial/parallel configuration what will assure a stable voltage/current while discharging all batteries evenly at the same time. One more measure to make the system more reliable was placing the circuitry box in an easily accessible area which will help in the case of diagnostic and repair.

To address the reliability issues within the robot software, the first action is to maintain a strong connection between the Jetson TK1 and the Arduino. Using USB serial communication will pass commands and power Arduino at the same time. Also, establishing a VNC connection can help in immediately addressing any problems within the TK1 and fix it by using a windows-based laptop. The VNC connection will be a SSH connection so that no one can interfere with the robot communication. Another issue was solved by using Arduino Mega, instead of the Jetson TK1, to establish the Pulse Width Modulation (PWM) of the robot as the PWM in Arduino Mega is more accurate and more reliable which will result in making commands and controls smoother with no lag, Appendix C is showing the updated code for the Arduino.

VII. Competition Results Summary

After several designs and modifications, the team was able to develop a new robot design that met the NASA RMC regulations and delivered the DustyTRON 2.0 robot to Kennedy space Center in Florida, shown in Fig. 19 and 20. The robot passed all inspections steps and competed against 56 other robots from all over United State and placed 16th. This experience was exceptional as the team members were able to implement all their engineering knowledge and skills to contribute to the Space exploration mission.

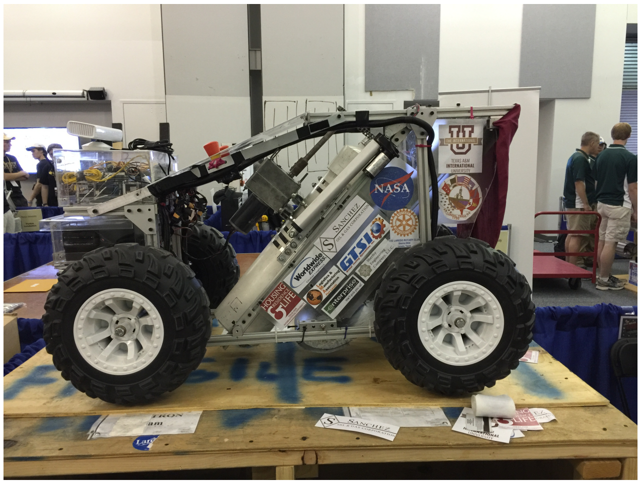


Fig. 19. DustyTRON 2 Robot - Final RMC 2016

VIII. Conclusion

In conclusion, DustyTRON team shown in Fig. 21 was created in August 2014, the team members were set to showcase their skills and utilize their knowledge in a state-of-art challenge. The NASA RMC was a great opportunity for them and placing the 16th out of 56 invited universities, was a respectable performance for a second-year team. The completed a new robot design that signifies the combination of mechanical, electrical, and computer engineering and computer science disciplines integrated into one cohesive system and present the power of their Systems Engineering background. The team worked on a completely new design being resourceful and utilize all the available sponsors and learned invaluable lessons about systems engineering principles and their implantation in a real-file problem solution.

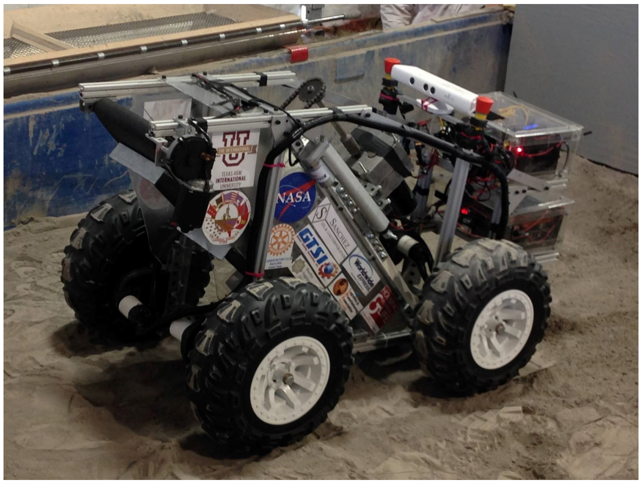


Fig. 20. DustyTRON 2 Robot - In Action

Fig. 21. DustyTRON 2 Team

The robot's locomotion system proved to be robust to handle the rough terrain that was encountered without failures, while staying under the expected weight and within a reasonable budget. One of the lesson learned from the project was the importance of time and team management, working to achieve all the goals taking into account the time and regulation constraint.

The team hopes to use and improve the robot for future competitions, therefore some of the improvements that can be tackled:

- Further develop the autonomy operation and enhance the robot vision software. The robot currently utilized partial autonomy to control its digging and dumping actions.
- Improve the sliding auger system, as it showed a significant success in digging and collecting regolith.
- Enhance the adaptive suspension system and improve the steering system, the robot has eclectically controlled suspension and drifting style steering system.

The team will continue to be involved with the local community outreach by presenting to local school to promote interest in the field of robotics, mainly NASA's current programs and projects in robotics. The team had presented to local middle and high school students their robotics projects to encourage them to join the STEM fields. Additionally, the team had been actively mentoring local FIRST Lego/Tech/Robotics teams in both stages of building and programming of their own robot, and host these competitions locally.

Acknowledgment

Thanks to NASA for providing such a great opportunity to participate in such competition. This project would not be successful without the support of Texas A&M International University (TAMIU) and our sponsors from Laredo, TX.

References

[1] Voosen, Paul. ”Mars rover steps up hunt for molecular signs of life.” (2017). Science 03 Feb 2017, Vol. 355, Issue 6324, pp. 444-445, DOI: 10.1126/science.355.6324.444

[2] Koris, Daniel R., and Jason Isaacs. (2017) ”A Formal Approach to Extended State Machines for Multi-Objective Robots Operating in Dynamic Environments.” Proceedings of the 2017 Midstates Conference on Undergraduate Research in Computer Science and Mathematics

[3] Tashtoush, T., Hernandez, R., Yanez, R., Gonzalez, J., Moreno, H., and Escobar, V. (2020). “Reverse-Twister Swarm Search Algorithm Design: NASA Swarmathon Competition”, International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), 7(1), pp.13-20.

[4] Hernandez, R., Yanez, R., Gonzalez, J., Moreno, H., Escobar, V., and Tashtoush. T., (2016) “Design of a Swarm Search Algorithm: DustySWARM Reverse-Twister Code for NASA Swarmathon.” Texas A&M International University, School of Engineering.

[5] Tashtoush, T., Gutierrez, O., Herrera, E., Medina, J., Peña, A., Varela, E., and Hernandez, R. (2020). “Design of a Swarm Search Algorithm: DustySWARM Spiral Epicycloidal Wave (SEW) Code for NASA Swarmathon”, International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), 7(1), pp.28-36.

[6] Gutierrez, O., Herrera, E., Medina, J., Peña, A., Varela, E., Hernandez, R., and Tashtoush. T. (2017) “Design of a Swarm Search Algorithm: DustySWARM Spiral Epicycloidal Wave (SEW) Code for NASA Swarmathon”. Texas A&M International University, School of Engineering.

[7] Tashtoush, T., Ruiz, C., Estevis, T., Herrera, E., Bernal, R., Martinez, R., and Reyna, L. (2020). “Square Spiral Search (SSS) Algorithm for Cooperative Robots: Mars Exploration”, International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), 7(1), pp.21-27.

[8] Tashtoush, T. H., Hernandez, R., Yanez, R., Gonzalez Jr, J., Moreno, H., & Escobar, V. (2020). Reverse-Twister Swarm Search Algorithm Design: NASA Swarmathon Competition. International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), 7(1), 13-20.

[9] Ruiz, C., Estevis, T., Herrera, E., Bernal, R., Martinez, R., Reyna, L., and Tashtoush, T., (2018) “Design of a Swarm Search Algorithm: Square Spiral Search (SSS) Algorithm for NASA Swarmathon”. Texas A&M International University, School of Engineering.

[10] Secor, P. (2016). “NASA Swarmathon”.

[11] Braccio, M. (2019). Design of a Robot for the 2019 NASA Robotic Mining Competition. In Proceedings of the Wisconsin Space Conference (Vol. 1, No. 1).

[12] Neubert, J. J. (2016). Using NASA's Robotic Mining Competition to Give Students a Quality Sys-tems Engineering Experience. In ASEE's 123rd Annual Conference & Exposition (pp. 4-11).

[13] Guerra,L., Murphy, G., and May. L., (2013). Applying Engineering to the Lunabotics Mining Competition Capstone Design Challenge. Proceeding of the ASEE Annual Conference and Exposition, June 2013.

[14] Stecklein, J. (2017, July). NASA's Robotic Mining Competition Provides Undergraduates Full Life Cycle Systems Engineering Experience. In INCOSE International Symposium (Vol. 27, No. 1, pp. 1456-1473).

[15] Chaput, A., 2016, ‘System Engineering Education for All Engineers - A Capstone Design Approach’. ASEE 123rd Annual Conference & Exposition, New Orleans, June 26-29, 2016.

[16] Mahmood, M. 2016, ‘Oakton Community College 2016 NASA Robotic Mining Competition Systems Engineering Paper’, paper presented to the 2016 NASA Robotic Mining Competition, Kennedy Space Center, Florida, 16-20 May.

[17] The University of Alabama in collaboration with Shelton State Community College, 2016, ‘Journey to Mars; 2016 Systems Engineering Paper’, paper presented to the 2016 NASA Robotic Mining Competition, Kennedy Space Center, Florida, 16-20 May.

[18] Charlotte 49er Miner Robotics, The University of North Carolina at Charlotte, 2016, ‘2016 Systems Engineering Paper’, paper presented to the 2016 NASA Robotic Mining Competition, Kennedy Space Center, Florida, 16-20 May.

[19] Illinois Robotics in Space (IRIS), University of Illinois at Urbana-Champaign, 2016, ‘Design and Development of the IRIS-6 Robotic Mining System’, paper presented to the 2016 NASA Robotic Mining Competition, Kennedy Space Center, Florida, 16-20 May.

[20] John Brown University Eaglenaut Robotics, John Brown University, 2015, ‘Robotic Regolith Excavation System’, paper presented to the 2015 NASA Robotic Mining Competition, Kennedy Space Center, Florida, 18-22 May.

[21] Chicago EDT Robotics, University of Illinois at Chicago, ‘Systems Engineering Report 2016, University of Illinois at Chicago, AMES-3 – Surus’, paper presented to the 2016 NASA Robotic Mining Competition, Kennedy Space Center, Florida, 16-20 May.

[22] Iowa State University Cyclone Space Mining, ‘2015-2016 Systems Engineering Paper’, paper presented to the 2016 NASA Robotic Mining Competition, Kennedy Space Center, Florida, 16-20 May.

[23] Dieter, G. E., & Schmidt, L. C. (2013). Engineering design. Boston: McGraw-Hill Higher Education.

[24] ”Rules and Rubrics”, Nasa.gov, 2016. [Online]. Available: http://www.nasa.gov/offices/education/centers/kennedy/technology/nasarmc/RulesRubricsResources

[25] Kapurch, S. J. (Ed.). (2010). NASA systems engineering handbook. Diane Publishing.

[26] Bellestri, S., Boil, T., Carswell III, M., et al. (2013). Alabama Lunabotic 2013 Systems Engineering Paper (Undergraduate Thesis) Retrieved from NASA.

[27] Alfaro, D., Aranguren, A., Duarte, T., De La Cruz H., Perez, G., Torres, A. Delgado, J., Melero, D., Vazquez, J. A., Charlton, B., Jose Guajardo, J., Flores, G., Garza, E., & Tashtoush, T. (2015). Systems Engineering Paper (Undergraduate Thesis) Retrieved from Texas A&M International University School of Engineering.

[28] Ay, N., Bertschinger, N., Der, R., Güttler, F., & Olbrich, E. (2008). Predictive information and explorative behavior of autonomous robots. The European Physical Journal B, 63(3), 329-339.

[29] Shotts Jr, W. E. (2012) “The Linux command line: a complete introduction” No Starch Press.

[30] Carson, E. M., Rivadeneira, J., Woodward, N. K., & Peterson, P. W. (2016). ”NASA Robotic Mining Competition 2015-2016”.

[31] Mueller, R. P. (2012) ”Lunabotics Mining Competition: Inspiration through Accomplishment” Thirteenth ASCE Aerospace Division Conference on Engineering, Science, Construction, and Operations in Challenging Environments, and the 5th NASA/ASCE Workshop On Granular Materials in Space Exploration.

[32] Williams, W. B., & Schaus, E. J. (2015). Design and Implementation of a Rocker-Bogie Suspension for a Mining Robot. In ASEE Southeast Section Conference.

[33] Liu, Y., Jeremy B., Zachary C., Jennifer B., John A.s, Madelyn D., David S., Cindy L. B., John B., and Christopher A.. ”Mechanical design, prototyping, and validation of a Martian robot mining system.” SAE International Journal of Passenger Cars-Mechanical Systems 10, no. 2017-01-1305 (2017): 177-182.

[34] Mueller, R., & Van Susante, P. (2011, September). A review of lunar regolith excavation robotic device prototypes. In AIAA SPACE 2011 Conference & Exposition (p. 7234).

Context Classification based on Mixing Ratio Estimation by Means of Inversion Theory

Kohei Arai
Faculty of Science and Engineering
Saga University, Saga City
Japan

***Abstract*—A contextual image classification method with a proportion estimation of the pixels composed of several classes, Mixed pixels (Mixels), is proposed. The method allows us to check the connectivity of separated road segments, which are observed frequently as discontinuity of roads in satellite remote sensing imagery. Under the assumption of almost same proportions for the Mixels in the discontinuous portion of road segments, a proportion estimation method utilizing Inverse Problem Solving is proposed. The experimental results with the simulation data including observation noise show 73.5~98.8(%) of improvements in terms of proportion estimation accuracy (Root Mean Square: RMS error), compared to the results from the previously proposed method with generalized inverse matrix. Also, usefulness of contextual classification based on the proposed proportion estimation was confirmed for the investigation of connectivity of roads in remotely sensed images from space.**

Keywords—Search engine; fuzzy expression; knowledge base system; membership function; mixed pixel: Mixel; context information; inverse problem solving

I. Introduction

Pixels constituting a satellite image are rarely composed of one ground covering class (Pure Pixe1) and are generally composed of a plurality of classes. These are called Mixel (Mixed Pixel), and the heterogeneous (Heterogeneous) area at the boundary between homogeneous (Homogeneous) areas is constituted by Mixel [1]. These are factors that reduce the image classification accuracy. It is not so difficult to classify Pure Pixel with high precision, and how to classify Mixel with high accuracy is a problem in image classification [2].

The author examined the context classification of Mixel in the heterogeneous area here. The spectral reflection characteristic of Mixel is considered as a mixture of the spectral reflection characteristics of multiple classes of Pure Pixel that constitute it and is given as a linear combination function weighted by the mixing ratio of each class. For this reason, in Mixel, if the spectral reflection characteristics of the class of Pure Pixel and the mixture ratio of Mixel that are constructed in any way from the image are known, the pixel is equivalent to being classified.

Mixel class mixing ratio estimation method based on the concept of feature mixing, using inverse matrix, generalized inverse matrix, quadratic programming [2]-[4], least squares method [5], surrounding 8 pixels. Considering the state of, the one obtained by linear regression has been proposed [6]. In addition, when calculating the spectral reflection characteristics of Pure Pixe1 from the image, the conventional method considers the residual of the observation vector and its estimated value in consideration of the scattering reflection characteristics of the observation target, fluctuation of observation conditions, and dispersion due to measurement error. While the solution was obtained under the condition of minimizing, a method of improving the estimation accuracy by applying the least square method to the class mixture ratio has been proposed [7].

The author is studying the expertly of image classification, focusing on road discontinuities that are scattered on satellite images [8], and trying to develop a knowledge base to check the connect ability of segments of discontinuous roads As the state of the discontinuous part and its surrounding pixels used in this knowledge base [9], the class mixture ratio that constitutes those pixels is taken up, the above method is improved and estimated [10], and it is used as contextual information. By using it, a method of checking road continuity is proposed.

In the following section, related research works and research background including motivation of the research are described. Then, the proposed context classification method is described followed by experimental method together with experimental results. After that, concluding remarks and some discussions are described.

II. Related Research Works

Classification by re-estimating statistical parameters based on auto-regressive model is proposed for purification of training samples [11]. Meanwhile, multi-temporal texture analysis in TM classification is proposed for high spatial resolution of optical sensor images [12]. On the other hand, Maximum Likelihood (MLH) TM classification considering pixel-to-pixel correlation is proposed [13].

Supervised TM classification with a purification of training samples is proposed [14] together with TM classification using local spectral variability is proposed [15]. A classification method with spatial spectral variability is also proposed [16] together with TM classification using local spectral variability [17].

Application of inversion theory for image analysis and classification is proposed [18]. Meanwhile, polarimetric SAR image classification with maximum curvature of the trajectory in eigen space domain on the polarization signature is proposed [19]. On the other hand, a hybrid supervised classification method for multi-dimensional images using color and textural features is proposed [20].

Polarimetric SAR image classification with high frequency component derived from wavelet multi resolution analysis: MRA is proposed [21]. Comparative study of polarimetric SAR classification methods including proposed method with maximum curvature of trajectory of backscattering cross section in ellipticity and orientation angle space is conducted and well reported [22].

Comparative study on discrimination methods for identifying dangerous red tide species based on wavelet utilized classification methods is conducted [23]. On the other hand, multi spectral image classification method with selection of independent spectral features through correlation analysis is proposed [24]. Image retrieval and classification method based on Euclidian distance between normalized features including wavelet descriptor is proposed [25].

On the other hand, wavelet Multi-Resolution Analysis: MRA and its application to polarimetric SAR classification is proposed [26]. Meanwhile, object classification using a deep convolutional neural network and its application to myoelectric hand control is proposed and evaluated its performances [27].

Image classification considering probability density function based on simplified beta distribution is conducted [28]. Also, Maximum Likelihood; MLH classification based on classified result of boundary mixed pixels for high spatial resolution of satellite images is proposed [29].

III. Research Background

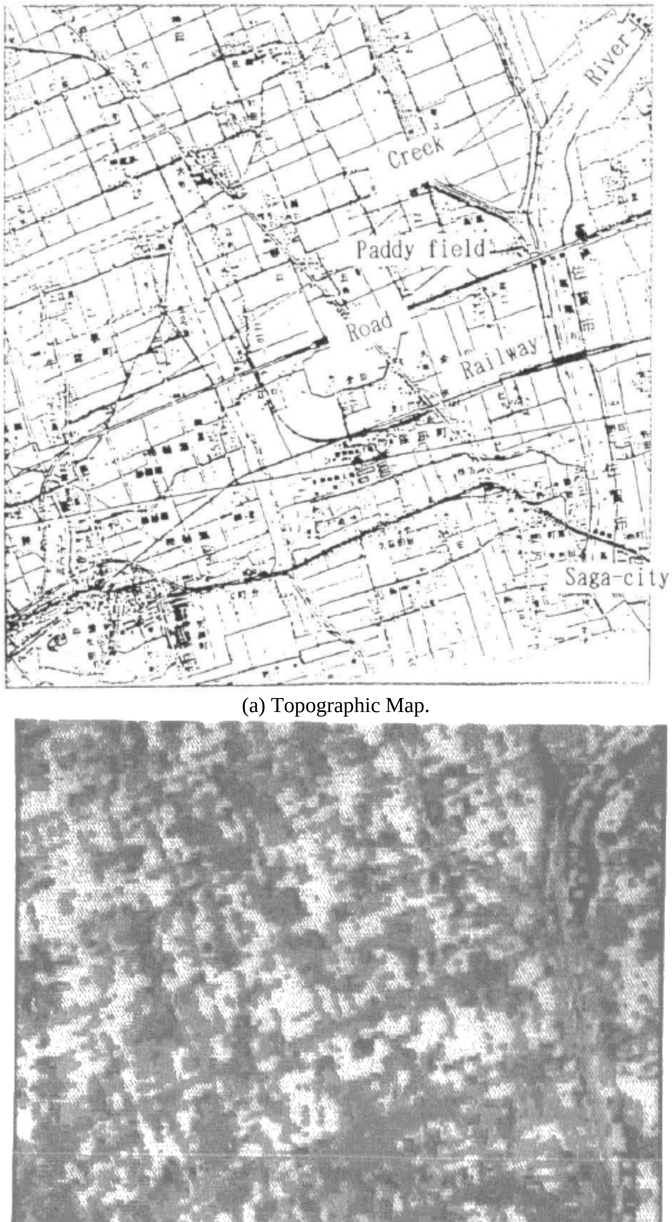
A. Motivation of the Research

Fig. 1(a) shows the topographic map of the intensive study area and Fig. 1(b) is a part of the image of the area including Saga city and Ushizu town acquired by LANDSAT-5 Thematic Mapper: TM in May 1986. This area is almost composed of paddy fields, as evident from the topographic map shown in the figure, and includes urban areas, residential areas, roads, railways, and rivers. Focusing on roads, there are road discontinuities determined by the relationship between the road width and the instantaneous field of view (IFOV) of the TM, the scanning direction and the angle between the roads, and the like.

The pixel of the discontinuous portion is located between the paddy fields in the periphery thereof in the spectral space and is rather close to the paddy field. Therefore, the pixel is classified into the paddy field using only the spectral information. The result of road extraction using only this spectrum information has discontinuous parts as shown in Fig. 2.

In order to examine whether these discontinuous pixels can be judged as roads, first, for these road segments extracted from the image, line likeness and directionality as texture information, which is one of spatial information, are used.

As a result, it was confirmed that it was a line element and the directionality was also consistent. However, in this study, which aims at automatic classification by a computer using knowledge base, it is not reliable to judge spectrally different pixels as the same road only by matching spatial information. For this reason, the class mixture ratio of pixels in the discontinuous part of the road and the class mixture ratio of surrounding pixels in the road segment are estimated as contextual information, and these are comprehensively determined to determine whether they are on the same road.



(a) Topographic Map.

(b) Landsat-5 TM Image.

Fig. 1. Portion of Landsat-5 TM Image of Saga City, Japan acquired in May 1986.

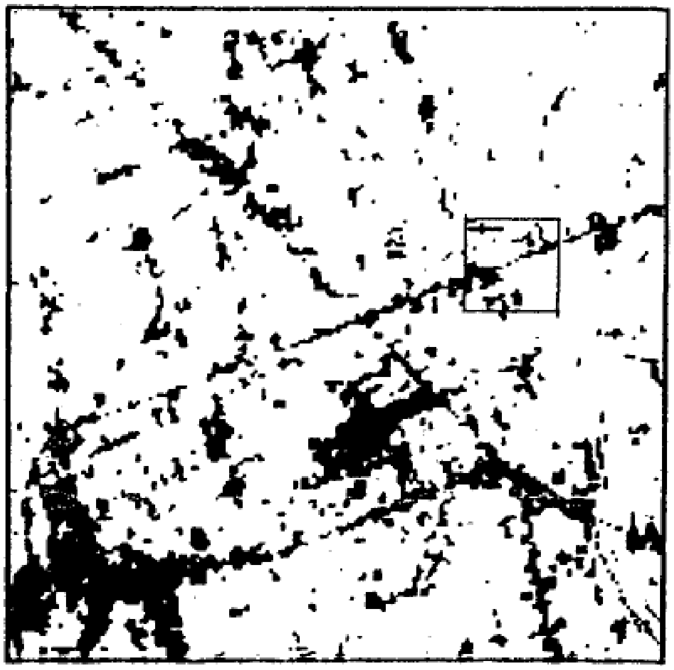


Fig. 2. An Example of Disconnected Portion of Road in Landsat-5 TM Image of Saga.

Specifically, Hough transform is applied to the image, road segments are re-extracted, and road segments at discontinuous portions are extracted, and pixels that appear to be roads are extracted by expanding the area with 3x3 windows around the extracted pixels. Then, their class mixture ratio was estimated. As an example of this, Fig. 3 shows the result of adapting a part of the above image.

B. *Class Mixing within Pixels using Conventional Generalized Inverse Matrix Estimation Method of Ratio*

Let be I is the M-dimensional observation vector of the pixel composed of *N* classes, *B* is the mixture ratio vector of those classes, and the spectral reflectance characteristics of the ground object corresponding to those classes, i.e., Pure Pixel in spectral space Assuming that the spectral vector of is *A*, it can be represented by the following linear combination function.

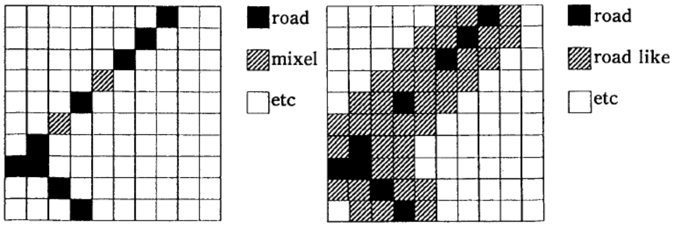
$$I=AB \tag{1}$$

$$I=(I_1, I_2, \dots, I_M)^t \tag{2}$$

$$A = \begin{bmatrix} A_{11} & \dots & A_{1N} \\ \dots & \dots & \dots \\ A_{M1} & \dots & A_{MN} \end{bmatrix} \tag{3}$$

$$B=(B_1, B_2, \dots, B_n)^t \tag{4}$$

This *I* is known because it is an observation vector. For *A*, it is known if a basic class is set in advance and their spectral reflectance characteristics are measured. However, the spectral reflection characteristics of the object measured on the ground are different from those of the same object appearing in the satellite image. In other words, the measured values are far from ground-based measured values due to the influence of the atmosphere and the ground surface.



(a) Road Pixels, Mixel and other. (b) Road, Road-Like Pixel and other.

Fig. 3. Extracted Road, Road-Like and the other Pixels.

In order to avoid this, the distribution of the basic class in the spectral space was obtained from the satellite image, and this average vector was determined as a representative of the spectral reflection characteristics of each class. Here, Eq. (6) is obtained by solving Eq. (1) under the condition that minimizes the norm between the observed value and the estimated value in Eq. (5), and the mixture of the set basic classes in Mixel is obtained.

$$\sqrt{\sum_{i=1}^{M} E_i^2} \rightarrow min., E = I - AB \tag{5}$$

$$B=(A^tA)^{-1}A^tI \tag{6}$$

IV. Proposed Method

A. *Estimation of Mixture Ratio of Pixel almost equal to Mixture Ratio of Preceding Pixel*

In the situation currently set, it is considered that the class mixture ratio between adjacent pixels of continuous boundary pixels may be substantially the same. For example, an asphalt road crossing a large paddy field as shown in Fig. 4 is a typical example. It is considered that the class mixture ratio of such continuous boundary pixels is often almost the same between adjacent pixels.

Therefore, for the pixel of interest, a constraint that minimizes the norm from the class mixture ratio of the boundary pixel adjacent to the pixel of interest is further added to Equation (5) to improve the estimation accuracy of the class mixture ratio of the pixel of interest. We propose a method to achieve. Here, considering not only the average value but also the variance, it results in a nonlinear optimization problem, but the solution in this case has already been proposed, so please refer to it.

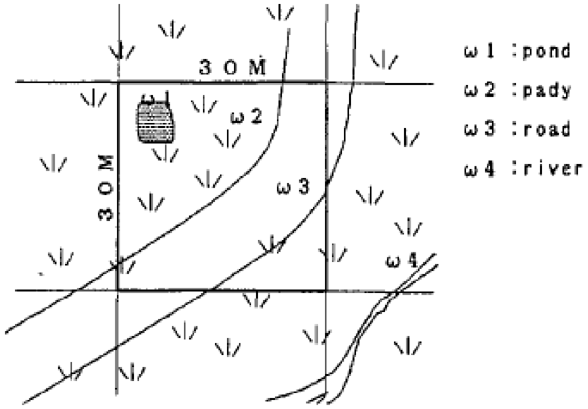


Fig. 4. An Example of Mixels, in the Discontinuous Portion of Road Segments, with almost Same Proportion r: Constraint Factor.

In addition, as pointed out in [7], the representativeness of the spectral reflectance characteristics of this basic class should be examined in consideration of physical observation conditions. However, in estimating the class mixture ratio proposed here, even if the mixture ratio is estimated assuming that a “road-like class” in which the pixel occupancy of the road is dominant is set from the beginning, It was judged that it was sufficient to judge the continuity of the road by estimating the mixture ratio of the discontinuous part of the existing road and the pixels around them, and it was decided not to depend on the representativeness.

If a vector representing the class mixture ratio of the adjacent boundary pixels is X, an expression representing the norm of the target pixel at the mixture ratio B can be written as follows.

$$\sum_{i=1}^{N}(B_i - X_i)^2 = (B - X)^t(B - X) \tag{7}$$

where

$$X = (X_1, X_2, \dots, X_N)^t \tag{8}$$

$$\sum_{i=1}^{N} B_i = 1 \tag{9}$$

$$B_i \geq 0 \tag{10}$$

X can be written as follows.

$$X = \begin{bmatrix} X_1\ (B_1 + B_2 + \cdots + B_N) \\ \cdot \\ \cdot \\ X_N(B_1 + B_2 + \cdots + B_N) \end{bmatrix} = \begin{bmatrix} X_1 & \dots & X_1 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ X_N & \dots & X_N \end{bmatrix} \begin{bmatrix} B_1 \\ \cdot \\ \cdot \\ B_N \end{bmatrix} \tag{11}$$

From Eq. (11),

$$(B - X) = \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} \begin{bmatrix} B_1 \\ \cdot \\ B_N \end{bmatrix} - \begin{bmatrix} X_1 & \dots & X_1 \\ \cdot & \cdot & \cdot \\ X_N & \dots & X_N \end{bmatrix} \begin{bmatrix} B_1 \\ \cdot \\ B_N \end{bmatrix}$$

$$= \begin{bmatrix} 1 - X_1 & -X_1 \dots & -X_1 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ -X_N & \dots - X_N & 1 - X_N \end{bmatrix} \begin{bmatrix} B_1 \\ \cdot \\ \cdot \\ B_N \end{bmatrix} \tag{12}$$

From the above, an expression representing the norm between the vector representing the class mixture ratio of the adjacent boundary pixels and the mixture ratio of the pixel of interest can be written as follows.

$$\sum_{i=1}^{N}(B_i - X_i)^2 = (YB)^t(YB) = B^tY^tYB \tag{13}$$

Eq. (1) is solved using this and the constraint condition (Eq. (14)) that minimizes the square error to obtain Eq. (16).

$$F = E^tE + B^tY^tYB \tag{14}$$

$$\frac{\partial F}{\partial B} = \frac{\partial\{(I - AB)^t(I - AB) + rB^tY^tYB\}}{\partial B} = rY^tYB - A^t(I - AB) = 0 \tag{15}$$

$$B = (A^tA + rY^tY)^{-1}A^tI \tag{16}$$

B. Determining the Strength of Constraints

In Eq. (16), if r is set to zero, it is equivalent to the method using the generalized inverse matrix, and conversely, if this is increased, the class mixture ratio vector of the pixel of interest approaches the class mixture ratio vector of the pixel preceding it without limit. Here, as a method for optimizing r, a method is considered in which the distance in the spectral space between adjacent boundary pixels is used as an index, and r is reduced if the distance is long, and is increased if the distance is short.

In other words, it is assumed that the pixel of interest for which the intra-pixel class mixture ratio is to be estimated is almost the same as the class mixture ratio estimated before it, so if some degree of similarity between pixels is known, it is used as an index. Then, the constraint condition can be changed. Although various similarities can be considered, the Euclidean distance in the spectral space is taken as an example here.

An example is shown below in Fig. 5 showing the relationship between the Euclidean distance in the spectral space between adjacent boundary pixels and the optimal r that minimizes the Root Mean Square: RMS error.

This figure plots the estimation accuracy of the class mixture ratio when the Euclidean distance between adjacent pixels is changed by 5 and the constraint condition factor r is changed by 2,000, and the thick solid line is the locus of the minimum value. In this example, the number of classes is set to 3, and their mixture ratio is set to 0.2, 0.0, and 0.8, respectively. In addition, the statistical noise between adjacent pixels is changed from the position in the spectral space to the center of gravity of the closed surface with the average value vector of each class as the vertex, and observation noise is reduced so that S / N becomes 10.

As is clear from Fig. 5, the estimation accuracy (Root Mean Square: RMS error) of class mixture ratio depends on r and Euclidean distance. In an extreme case, if the Euclidean distance between adjacent pixels is 0, a large constraint condition is applied, and the estimation error can be made as close to 0 as possible. Conversely, when the Euclidean distance is large, the estimation accuracy is better if no constraint condition is applied. That is, it becomes equivalent to the mixture ratio estimation itself by the generalized inverse matrix.

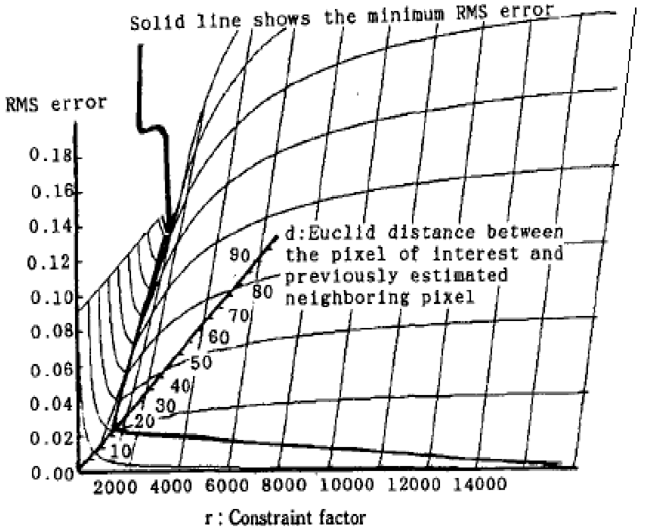


Fig. 5. Relationship between Optimum r and Statistical Distance between the Pixel of Interest and Previously Estimated Neighboring Boundary Pixels (Mixels) in Spectral Feature Space.

C. Class Settings and their Number

In general, in estimating the class mixture ratio, the larger the number of classes, the lower the estimation accuracy of the mixture ratio. Therefore, we considered limiting the number of classes as much as possible. Here, the pixels at the discontinuous portion of the road of interest are boundary pixels that are spectrally heterogeneous and exist at the boundaries of different class regions. The image has spatial information based on the macroscopic structure (for example, a network of roads, railways, rivers, etc.), which can be extracted by an edge operator or the like. Furthermore, it is confirmed that the boundary pixels are spectrally boundary by the homogeneity test [1] and the like.

There are several types of pixels of the set class around the boundary pixel (at most, the number of pixels around eight pixels around the boundary class). When estimating the class mixture ratio based on Eq. (16), the smaller the number of set classes, that is, the smaller the number of unknowns, the better the estimation accuracy, but only for the boundary pixels such as roads. Then the number of classes will not be so large. When determining the number of classes to be set, the frequency distribution of the number of classes of eight neighboring pixels around the boundary pixel may be checked and determined in advance.

V. Experiments

A. Simulation Data

Set the number of classes to 3, and Mixel was simulated as a parameter. The generated random number is signaled so that the S / N ratio becomes 10. Table I shows a comparison among the proportion estimation accuracies derived from Generalized Inverse Matrix Method (GIMM), the Least Square Method with a Prior Proportion (LESP) in case of which the Euclidian distance between the pixel of interest and the previously estimated boundary pixel is 3.

Since the proposed method is based on the generalized inverse matrix method and adds constraints, the simulation results are used to estimate the generalized inverse matrix and the estimation accuracy and RMS error of the method proposed in this paper. As a result, as shown in Table I, the proposed method showed that the RMS error in estimating the mixture ratio was 73.5-98.8% better than that using the generalized inverse matrix.

B. Actual Data

The data of the area including Saga city and Ushizu town acquired in May 1986 by LANDSAT-5 / TM was used. The area is composed of paddy fields, including urban, residential, road, railway and river networks. This is called Saga data here.

Focusing on the area surrounded by the solid line in Fig. 2, three pixels are selected from each segment of the paddy field located away from the road, the paddy field on the side of the road, the road, and the unconnected part, and their mixture ratio was estimated. Fig. 6 shows the selected pixel position and Table II shows the estimation results.

The results showed that the discontinuous pixels had a large vegetation ratio. When we went to the corresponding place and investigated the situation, we confirmed that the branches and leaves of the trees on the road covered the road.

The matrix A representing the spectral reflectance characteristics of each class in Eq. (15) was obtained by extracting pixels considered to be representative of the class from the image and averaging them. However, since the extracted pixels themselves are often Mixel, an estimated from these is poor in the representativeness of each class of paddy field, road, and vegetation.

In particular, the pixels corresponding to the road class are greatly affected by the ground covering having spectral reflectance characteristics other than roads, and do not represent the true road spectral reflectance characteristics. However, even if the mixture ratio is estimated assuming that a “road-like class” in which the pixel occupancy is dominant is set from the beginning, the discontinuous portion of the road under consideration and their mixture ratio estimation Is considered to be sufficient for the determination of road continuity according to.

TABLE I. A Comparison among the Proportion Estimation Accuracies Derived from Generalized Inverse Matrix Method (GIMM), the Least Square Method with a Prior Proportion (LESP) in Case of which the Euclidian Distance between the Pixel of Interest and the Previously Estimated Boundary Pixel is 3

Mixing Ratio			GIMM			LESP		
Road	Paddy	Water	Road	Paddy	Water	Road	Paddy	Water
0.5	0.5	0	0.507	0.417	0.076	0.499	0.501	0
RMSE			0.0646			0.0008		
Optimum r			N/A			580		
0.5	0	0.5	0.589	0.006	0.405	0.497	0.026	0.477
RMSE			0.0756			0.02		
Optimum r			N/A			440		
0.2	0.8	0	0.252	0.672	0.076	0.214	0.786	0
RMSE			0.0911			0.0118		
Optimum r			N/A			1000		
0.4	0.4	0.2	0.497	0.393	0.11	0.397	0.405	0.198
RMSE			0.0765			0.0032		
Optimum r			N/A			580		

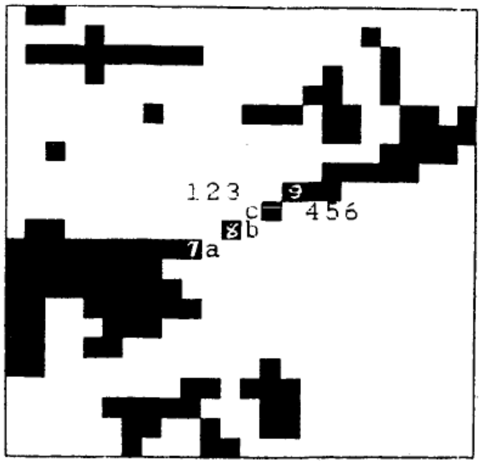


Fig. 6. Designated Pixels for Investigation of Proportion Estimation.

TABLE II. ESTIMATED MIXING RATIOS OF ROAD, PADDY FIELD AND WATERBODY CLASSES FOR THE EXTRACTED PIXELS

No.	Mixing Ratio			Remarks
	Road	Paddy	Water	
1	0.198	0.612	0.19	Paddy field apart from road
2	0.193	0.674	0.133	
3	0.145	0.622	0.243	
4	0.287	0.612	0.101	Paddy field beside road
5	0.034	0.582	0.384	
6	0.078	0.552	0.37	
7	0.51	0.399	0.001	Road
8	0.484	0.395	0.121	
9	0.637	0.231	0.132	
a	0.333	0.536	0.131	Disconnected segments of road
b	0.343	0.517	0.14	
c	0.443	0.377	0.18	

VI. CONCLUSION

The method of estimating the class mixture ratio in adjacent boundary pixels proposed in this paper is more effective than the conventional method using the generalized inverse matrix when the class mixture ratio of successive boundary pixels is almost equal. The estimation accuracy is excellent. That is, if the Euclidean distance between consecutive boundary pixels is 3 and the S / N due to observation noise is 10, the RMS value of the estimation error of the proposed method is reduced to about 1/4 to 1/81 of that of the generalized inverse matrix.

It was confirmed that it could be reduced. This method is effective especially when the Euclidean distance between consecutive boundary pixels is short, such as a road running in a large paddy field or in an urban area, that is, when the class mixture ratio of the boundary pixels is about a bit. If this precondition does not hold and the distance between consecutive boundary pixels is long, it is equivalent to a generalized inverse matrix.

This estimation method can be applied not only to one of the knowledge bases using context information such as "trees wear on the road" as an example, but also to more knowledge bases.

VII. FUTURE RESEARCH WORKS

The proposed method is adopted in the real earth observation satellite imagery data, and it is a future subject to realize a more usable context-based classification method.

ACKNOWLEDGMENT

The author, also, would like to thank Professor Dr. Hiroshi Okumura and Professor Dr. Osamu Fukuda for their valuable discussions.

REFERENCES

[1] Ketting, R.L. and D.A. Landgrebe, Classification of multispectral imabe data by Extraction and Classification of Homogeneous Objects, Proceedings of the 1975 Machine Processing of Remotely Sensed Data Symposium, pp. 2 A-1-2A-11, Purdue Univ., USA, 1975.

[2] Horwitz, H,M., et al., Estrimating the Proportions of objects within a single resolution element of a multispectral scanner, Proceedings of the 7 th Int. Symp. on Remote Sensing of Environment, No. 10259-1-X, pp. 1307-1320, 1971.

[3] Nalepka, R.F., et al., Estimating proportions of objects from multispectral data, NASA Report No. CR-WRL 31650-73-T, 1972.

[4] Hall, F.G., Satellite Remote Sesning: An integral tool in aquiring global crop production information, Proceedings of the 1982 Machine Processing of Remotely Sensed Data Symposium, 10-22, 1982.

[5] Minoru Inamura, Analysis of Remote Sensing Image Data Based on Category Decomposition, IEICE Journal, Vol.J70-C, no.2, pp.241-250, 1987.

[6] Rikimaru, Kamijo, Oshima, Development of Simple Estimation Method for In-pixel Spectral Information, Journal of the Japan Society of Photogrammetry, Vol. 27, no. 6, pp. 23-34, 1988.

[7] Ito, Fujimura, Area ratio estimation by category decomposition of pixels, Transactions of the Society of Instrument and Control Engineers, Vol.23, No.8, pp.20-25, 1987.

[8] Kohei Arai, Terayama, Teramoto, Matsumoto, Fujikyu, Tsuchiya, Image Classification by Category Decomposition (I)-Context Classification with Category Decomposition by Inverse Problem Solving I, The Remote Sensing Society of Japan, 119-122, 1990.

[9] Ozaki, Taniguchi, Image Processing, Kyoritsu Shuppan, 1983.

[10] Matsumoto, Fujikyu, Tsuchiya, Arai, Category decomposition based on maximum likelihood estimation, Journal of the Japan Society of Photogrammetry, Vol. 30, No. 2, pp. 25-34, 1991.

[11] Kohei Arai, Classification by Re-Estimating Statistical Parameters Based on Auto-Regressive Model, Canadian Journal of Remote Sensing, Vol.16, No.3, pp.42-47, Jul.1990.

[12] Kohei Arai, Multi-Temporal Texture Analysis in TM Classification, Canadian Journal of Remote Sensing, Vol.17, No.3, pp.263-270, Jul.1991.

[13] Kohei Arai, Maximum Likelihood TM Classification Taking into account Pixel-to-Pixel Correlation, Journal of International GEOCARTO, Vol.7, pp.33-39, Jun.1992.

[14] Kohei Arai, A Supervised TM Classification with a Purification of Training Samples, International Journal of Remote Sensing, Vol.13, No.11, pp.2039-2049, Aug.1992.

[15] Kohei Arai, TM Classification Using Local Spectral Variability, Journal of International GEOCARTO, Vol.7, No.4, pp.1-9, Oct.1992.

[16] Kohei Arai, A Classification Method with Spatial Spectral Variability, International Journal of Remote Sensing, Vol.13, No.12, pp.699-709, Oct.1992.

[17] Kohei Arai, TM Classification Using Local Spectral Variability, International Journal of Remote Sensing, Vol.14, No.4, pp.699-709, 1993.

[18] Kohei Arai, Application of Inversion Theory for Image Analysis and Classification, Advances in Space Research, Vol.21, 3, 429-432, 1998.

[19] Kohei Arai and J.Wang, Polarimetric SAR image classification with maximum curvature of the trajectory in eigen space domain on the

polarization signature, Advances in Space Research, 39, 1, 149-154, 2007.

[20] Hiroshi Okumura, Makoto Yamaura and Kohei Arai, A hybrid supervised classification method for multi-dimensional images using color and textural features, Journal of the Japanese Society of Image Electronics Engineering, 38, 6, 872-882, 2009.

[21] Kohei Arai, Polarimetric SAR image classification with high frequency component derived from wavelet multi resolution analysis: MRA, International Journal of Advanced Computer Science and Applications, 2, 9, 37-42, 2011.

[22] Kohei Arai Comparative study of polarimetric SAR classification methods including proposed method with maximum curvature of trajectory of backscattering cross section in ellipticity and orientation angle space, International Journal of Research and Reviews on Computer Science, 2, 4, 1005-1009, 2011.

[23] Kohei Arai, Comparative study on discrimination methods for identifying dangerous red tide species based on wavelet utilized classification methods, International Journal of Advanced Computer Science and Applications, 4, 1, 95-102, 2013.

[24] Kohei Arai, Multi spectral image classification method with selection of independent spectral features through correlation analysis, International Journal of Advanced Research in Artificial Intelligence, 2, 8, 21-27, 2013.

[25] Kohei Arai, Image retrieval and classification method based on Euclidian distance between normalized features including wavelet descriptor, International Journal of Advanced Research in Artificial Intelligence, 2, 10, 19-25, 2013.

[26] Kohei Arai, Wavelet Multi-Resolution Analysis and Its Application to Polarimetric SAR Classification, Proceeding of the SAI Computing Conference 2016.

[27] Yoshinori Bando, Nan Bu, Osamu Fukuda, Hiroshi Okumura, Kohei Arai, Object classification using a deep convolutional neural network and its application to myoelectric hand control, Proceedings of the International Symposium on Artificial Life and Robotics (AROB2017), GS12, 2017.

[28] Kohei Arai, Image classification considering probability density function based on Simplified beta distribution, International Journal of Advanced Computer Science and Applications IJACSA, 11, 4, 481-486, 2020.

[29] Kohei Arai, Maximum Likelihood Classification based on Classified Result of Boundary Mixed Pixels for High Spatial Resolution of Satellite Images, International Journal of Advanced Computer Science and Applications, Vol. 11, No. 9, 24-30, 2020.

AUTHOR'S PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is a Science Council of Japan Special Member since 2012. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 55 books and published 620 journal papers as well as 450 conference papers. He received 66 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.html.

Concurrent Detection of Linear and Angular Motion using a Single-Mass 6-axis Piezoelectric IMU

Hela Almabrouk[1]
Université de Sfax
Ecole Nationale d’Ingénieurs de Sfax, Sfax Tunisia
Faculté des Sciences de Monastir
Université de Monastir, Monastir Tunisia
C2N, UMR 9001, Université Paris-Saclay
91120 Palaiseau, France

Mohamed Hadj Said[2]
Center for Research in Microelectronics and Nanotechnology (CRMN)
Sousse Technopole
Novation City, Tunisia

Fares Tounsi[3], Brahim Mezghani[4]
Université de Sfax, Ecole Nationale d’Ingénieurs de Sfax
UR-METS, Sfax, Tunisia

Guillaume Agnus[5]
Centre de Nanosciences et de Nanotechnologies (C2N)
UMR 9001, Université Paris-Saclay
91120, Palaiseau, France

Yves Bernard[6]
GeePs | Group of Electrical Engineering - Paris, CNRS,
CentraleSupélec, Université Paris-Saclay, Sorbonne
Université, 3 & 11 rue Joliot-Curie, Plateau de Moulon,
91192, Gif-sur-Yvette CEDEX, France

***Abstract*—This paper exhibits operating system and performances of a novel single-mass 6-axis Inertial Measurement Unit (IMU) using piezoelectric detection. The electronic processing circuitry for the concurrent detection of linear and angular motion is proposed. The IMU structure is based on the use of 2 rings, connected with eight electrodes, implemented on the top of a piezoelectric membrane used for both sense and drive modes. The four inner electrodes are used for components detection due to the direct piezoelectric effect, while the outer electrodes are used to generate the drive mode due to the reverse piezoelectric effect. Through finite element analysis, we show that linear accelerations generate an offset voltage on the sensing electrodes, while angular rates lead to a change in the amplitude of the initial AC signal caused by the drive mode. The present work represents an innovative design able to separate 6 motion data from signals using only 4 electrodes. The specific electronic circuitry for acceleration and angular rate data dissociation shows a very efficient method for signal separation since no leakage readout occurs in all six axes. Besides, other particular interest is that under no circumstances, angular outputs disturb or affect acceleration ones and vice versa. The evaluated sensitivities are 364 mV/g and 65.5 mV/g for in-plane and out-of-plane linear accelerations, respectively. Similarly, angular rates sensitivities are 2.59 mV/rad/s and 522 mV/rad/s.**

Keywords*—*Inertial measurement unit; piezoelectric detection; angular rate; linear acceleration; electronic circuitry

I. Introduction

Inertial sensors, including magnetometers, gyroscopes and accelerometers are widely demanded in various fields such as navigation, automotive industry, robotic and military domains [1-4]. Commonly, each of these sensors has typically three degrees of freedom to measure from the three-dimensional surrounding space. Inertial measurement units (IMU) integrate multiple miniature inertial sensors to obtain comprehensive inertial parameters of the moving object, including the attitude, position, and speed information [5]. This information comes from the measure of rotational and linear inertial data such as angular velocity and gravitational force. IMU devices are highly independent and less susceptible to outside interference with fast data updates and strong stability. These devices are groundbreaking sensors that have been extensively studied by various researchers [6-9]. The multi-axis motion detection mechanisms rely on various sensing types. It can be piezoelectric [10-11], thermal [12], capacitive [13], piezoresistive [14], etc. However, piezoelectric inertial sensors are being known to offer important advantages compared to their counterparts. In fact, they can be used for wide range of frequencies by offering a very large dynamic range. Indeed, this type of inertial sensor is suitable for low amplitude shock measurements providing a very low noise level [15]. Moreover, piezoelectric sensors are known by offering reduced power consumption.

Many recent inertial sensors design allow to achieve linear and/or angular accelerations measurements through a single-core detecting element [16-18]. Tuning fork and disk resonating structures are among the most successful single-core devices [10, 16, 17]. The tuning fork resonator relies on the piezoelectric effect for the drive mode excitation and uses either piezoelectric or capacitive effect for the detection. The tuning fork structure is known by its simplicity and easy assembly with the associated electronic detection circuit. Nevertheless, it presents a low sensitivity and uses indirect method to measure the deflection especially the capacitance detection [16]. The other structure is composed by a resonant slotted disk supported by a central cylindrical post and surrounded by capacitive electrodes used to force and sense vibration. The resonator disk has the advantages of providing high quality factor and thermal stability due to the low anchor loss, air damping effect (perfored disk) and thermo-elastic dissipation [19]. Moreover, the symmetrical propriety of the disk structure offers a high sensitivity [17, 20]. In addition, the resonant disk devotes a large area of multiple drive and sense electrodes so that it becomes easier to operate and measure signals, using this structure, compared to other inertial sensors [21].

Inertial measurement unit devices are widely used in various applications due to the important information that it can provide including acceleration and angular rate data of a movable object. However, because of the use of two separated blocks of sensors, one for accelerometer and other for angular rate sensor, several limitations bound IMU development which consist mainly in size, cost and energy consumption. Therefore, IMUs are being restricted to relative bulk applications and thus, although its importance, conventional IMU devices become unsuitable to consumer applications unless reducing devices size. In this frame, along this paper, we detail the design performances of a 6-axis single-mass piezoelectric inertial measurement unit using one single sensor instead of two. Considering the significant advantages related to resonant structures, we are proposing an emerging mesoscale piezoelectric 6-DOF design capable to simultaneously detect 3-axis angular rate and 3-axis acceleration components, using a single-core detecting element.

The proposed structure uses a cylindrical proof mass as an oscillator, attached to the bottom center of an anchored piezoelectric disk. This configuration can improve performances due to the use of a great modal effective mass compared to other resonator designs [10, 21]. The structure uses piezoelectric effect to drive and sense the 6-DOF (Degree of Freedom) which is an advantage compared to the tuning fork and disk resonator structures, which generally detects only angular motion components [17, 19]. Moreover, the present design employs only four electrodes for driving the seismic mass and four other ones for separating the 6-DOF outputs. Finite Element Modeling (FEM) is used to simulate the detection mechanism performances of both angular velocity and linear acceleration on our proposed piezoelectric design and we extract output signals of each inertial component are extracted. FEM simulations are performed using the COMSOL® Multiphysics software package. Moreover, we are interested in the design of the conditioning part of the detection system using Orcad-PSPICE software.

In this paper, a presentation of the studied IMU design and its operation system will be firstly presented before going through an analytical analysis of the system operation by building the drive and sense mode equations. Then, numerical investigation of both linear and angular rate detections will be detailed. Results in terms of sensitivity will be compared with other designs from literature. We are also interested in the presentation of the separation electronic circuitry from which signals relative to three-axis acceleration and three-axis angular rate inputs will be faithfully identified.

The main objective of this work is to design a 6-axis sensor with the corresponding electronic circuit able to distinguish acceleration data from angular velocity ones, from the same sensing electrodes. We consider signals extracted from FEM simulations as input data for the development of the electronic circuitry.

II. Presentation of the IMU Design and its System's Operation

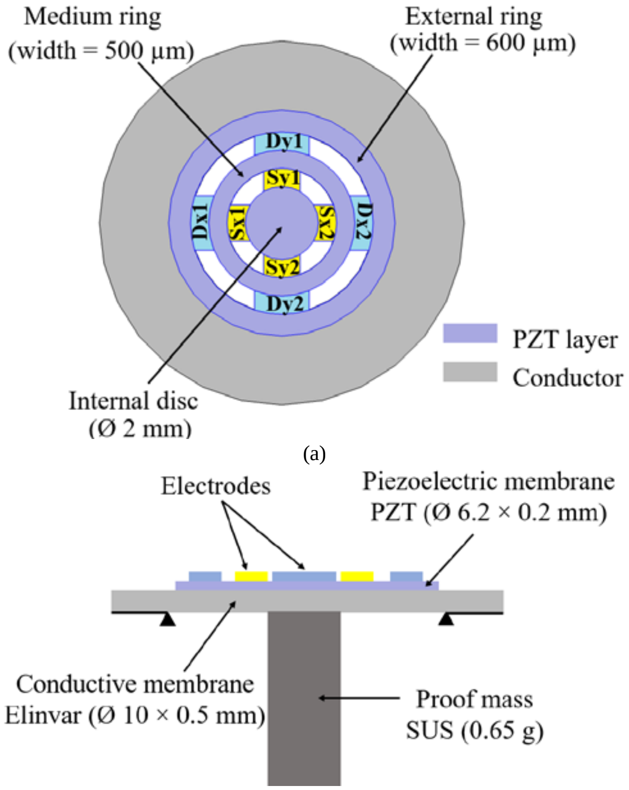
The main advantage of the novel developed IMU sensor is the simultaneous detection of the 6-DOF inertial motions using a single seismic mass-based structure. The purpose and the challenge of the present work is the analysis of the mechanical IMU structure motion along with its corresponding electronic circuitry capable to separate acceleration from angular velocity data.

The proposed geometry of the 6-axis motion sensor is the millimeter range. The sensitive element, i.e. the piezoelectric membrane, is made of PZT material. It is basically formed by an external ring, a medium ring and an internal disk which are interconnected using attachments forming the sensitive part of the IMU design (Fig. 1a). The used structure is inspired from a 6-DoF IMU design proposed by Okada et al. [10] and it has been proved through previous works that our new proposed IMU design offers better performances [22, 23].

The shape of the flexible piezoelectric membrane has been carefully designed to generate maximum stress profiles in the sensitive parts, on which electrodes are placed, when subjected to any motions (linear accelerations and/or angular rate rotations). This PZT layer is superimposed on a circular diaphragm made of high precision Elinvar which is an alloy that has low thermal expansivity and an elasticity modulus virtually unaffected by changes in temperature. A proof mass is used as an oscillator and is attached to the center of the bottom conductive membrane (Fig. 1b).

The operating principle is based on two essential modes; drive and sense modes. Drive mode is used to bring the oscillator, i.e. the proof mass, into a rotational motion in the *x/y* plane by applying an AC voltage with constant amplitude on drive electrodes (D_{x1}, D_{x2}, D_{y1} and D_{y2}) and sense mode is used to readout and measure the output signals from sense electrodes (S_{x1}, S_{x2}, S_{y1} and S_{y2}). These drive and sense electrodes have a width of 500 µm and are implemented on the top of the piezoelectric membrane. Their locations have been previously analyzed so that sense electrodes are implemented at maximum induced stress positions [24]. Attachments where sense electrodes are implemented, are designed to be tinier than those of drive electrodes in order to avoid losses and collect the totality of charges for more sensitivity.

The linear acceleration detection is accomplished based principally on exploiting the direct piezoelectric effect (Fig. 2). In fact, the applied acceleration is converted into a mechanical force given by Newton's second law (F=m.a). Depending on its direction, the seismic mass will tilt contrarily producing the maximum of deformation in the sensing electrodes location. Thanks to the direct piezoelectric effect, induced stress on the PZT membrane will be converted into an output voltage proportional to the acceleration input. On the other hand, angular rate detection employs the Coriolis effect phenomenon. Indeed, when the sensor is subjected to an angular motion, a deviation in the motion path of the oscillating mass arises due to the induced Coriolis force. Thus, angular velocity measurement is fundamentally related to the presence of a drive mode (primary permanent vibrating) and a second mode called sense mode. Generated Coriolis force will be orthogonal to both drive motion direction and rotation input axis. Therefore, mass vibration amplitudes in the sense mode are then proportional to the applied angular rates and coupled to the drive vibration. Since the main preeminence of the sensor is its ability to detect 6-DOF simultaneously, so the block diagram in Fig. 2 combines the two detection principles previously described.



(a)

(b)

Fig. 1. Proposed Design of the New 6-Axis Inertial Sensor based on a Single Proof-Mass (a) Top view Showing Drive and Sense Electrodes Arrangement, and (b) Cross Sectional view.

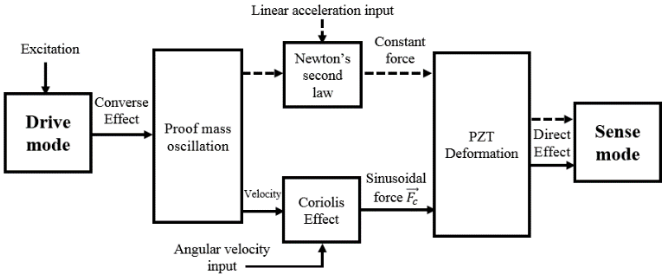


Fig. 2. Block Diagram Describing the Simultaneous Sensing of Linear and Angular Motion Principle (Acceleration Detection is based on Newton's Second Law, While Angular Rate Detection is based Coriolis Effect).

III. Analytical Analysis of the IMU Design

In this section we are introducing an analytical representation of the operation principle of the IMU design through the establishment of drive and sense modes equations.

Indeed, as an angular rate sensor, for each axis detection, the motion equation of the resonant sensor structure can be described using a mass-spring-damping system with two degrees of freedom as illustrated in Fig. 3. Each mode (Drive & Sense) is then represented by a mass, m, a spring and a damper. For each mode corresponds a stiffness k and a damping factor η (Fig. 3). An analytical description of the motion amplitude could be obtained by developing relations between the model and system variables through ordinary differential equations.

A. Drive Mode Equation

Based on the 1st fundamental principle of dynamics, projection on drive mode direction (similarly for sense mode), allows us to write:

$$m\ddot{x}_d = F_d - \eta_d\dot{x}_d - k_d x_d \tag{1}$$

where F_d is the excitation force and x_d is the corresponding displacement.

When considering ω (ω_d^2=$\sqrt{k_d/m}$), and Q (Q_d=ω_d m/η_d) that are, respectively, the natural frequency and the quality factor of the two modes, "(1)" becomes as follows:

$$\frac{F_d}{m} = \ddot{x}_d + \frac{\omega_d}{Q_d}\dot{x}_d + \omega_d^2 x_d \tag{2}$$

By applying Laplace transform, it is possible to extract drive mode displacement given by:

$$x_d = -j\frac{Q_d}{m\omega_d^2}F_d \tag{3}$$

We notice that the drive vibration is in phase quadrature (90° phase shift) with the excitation force and the drive vibration amplitude is presented as follows:

$$X_d = \frac{Q_d}{m\omega_d^2}F_d \tag{4}$$

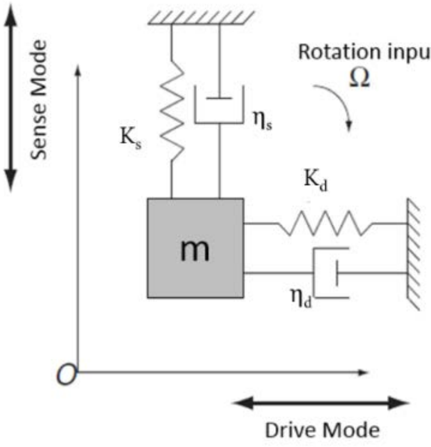


Fig. 3. Modeling of the Vibrating Sensitive Element Operation: Each Mode (Drive and Sense) is represented by a Mass-Spring-Damping System [25].

B. *Sense Mode Equation*

As explained before, under drive oscillation, sense mode is generated through the appearance of the Coriolis Effect along a perpendicular axis to both drive mode and the angular rate input direction. Thus, the equation of movement in the sense mode will be:

$$m\ddot{x}_s = -\eta_s\dot{x}_s - k_s x_s + 2m\Omega\dot{x}_d \tag{5}$$

When inserting the natural frequency and the quality factor expressions relative to sense mode vibration, the sense mode equation becomes:

$$2\Omega\dot{x}_d = \ddot{x}_s + \frac{\omega_s}{Q_s}\dot{x}_s + \omega_s^2 x_s \tag{6}$$

After passing to frequency domain, the ratio between the sense displacement x_s and the input rotation, Ω is expressed as:

$$\frac{x_s}{\Omega} = \frac{2j\omega_d x_d}{(\omega_s^2-\omega_d^2)+j\frac{(\omega_s\omega_d)}{Q_s}} \tag{7}$$

We consider that drive and sense resonant frequencies of the sensitive element are different (ω_s-ω_d $\neq$ 0) and the sense mode quality factor is important. Therefore, the expression of the sense vibration amplitude in function of the drive displacement amplitude is given by:

$$x_s = j\frac{2\omega_d\Omega}{\omega_s^2-\omega_d^2}x_d \tag{8}$$

From “(8)”, we can conclude that the vibration of the drive and sense mode are in quadrature. When considering that sense and drive frequencies are too close to each other but not equal, $\omega_s - \omega_d$ will be negligible compared to ω_s. Thus, “(8)” can be concluded as follow:

$$\frac{X_s}{\Omega} = \frac{X_d}{|\omega_s-\omega_d|} \tag{9}$$

The expression exhibited in “(9)” presents the mechanical scale factor of the studied gyro-part of the 6-DOF motion sensor expressed in (m/°/s) in terms of the drive mode vibration amplitude. Using the same theory, the dynamic response of the accelerometer is obtained in the frequency domain as:

$$\frac{|X|}{F} = \frac{1}{m\sqrt{(\omega_s^2-\omega_d^2)^2+4\varepsilon^2\omega_s^2\omega_d^2}}\ ;\ \varepsilon = \frac{\eta}{2\sqrt{km}} \tag{10}$$

where F is the applied force and ε is the damping ratio term.

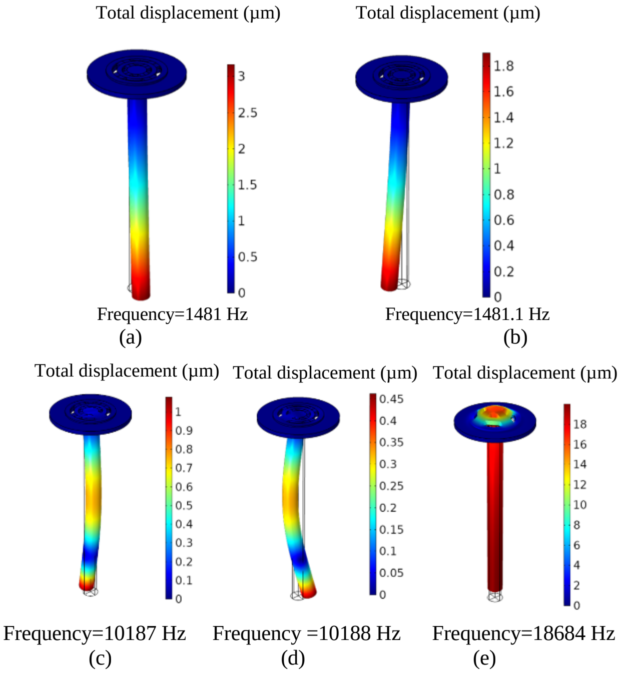
From “(9)”, we can conclude that in order to maximize the scale factor of the studied motion sensor we have to maximize the drive vibration amplitude. However, this is generally limited by the maximum stress allowed by the material to avoid structural damages. In addition, we can notice that the frequency difference between drive and sense modes must be minimized. Thus, we can conclude that in order to increase the sensitivity of the sensor, the sensitive element must be dimensioned with a manner that the resonance frequencies of the two modes be very close to each other.

The illustration of analytical equations that define the sensor operation system has a great interest of understanding the theories behind the sensor operation when dealing with FEM simulations on a one hand. On another hand, this analytical approach gives a deep insight to define a set of principal parameters that influence the design. This will considerably help to optimize the sensor performances that will be exhibited in the next section.

IV. Numerical Investigation of the 6-Axis IMU Design

A. *Drive Mode Generation*

Drive mode consists in making the structure oscillate into a circular trajectory in the *x/y* plane. This mode is chosen in such a way that the proof mass undergoes an oscillation motion in both *x*-axis and *y*-axis directions. This is achieved by applying four AC voltage signals, having a phase shift of 90 deg, to the four drive electrodes. The choice of the drive frequency is based on an Eigen frequency study that should be performed firstly through a finite element simulation to figure out the different structure normal modes. The two first in-plane resonant frequencies were found to be 1481 Hz and 1481.1 Hz, respectively along *x*- and *y*-axes as shown in Figs. 4a and 4b. Indeed, symmetrical flexure is obviously present along both *x*- and *y*-axis which produces the same modal frequency of the design. The next modes, shown in Fig. 4c and 4d, are non-axisymmetric. The next out-of-plane mode along *z*-axis, was found to be around 18 kHz (Fig. 4e). Mode 1 will be the one used for successfully completing the proposed configuration.



Frequency=1481 Hz
(a)

Frequency=1481.1 Hz
(b)

Frequency=10187 Hz
(c)

Frequency =10188 Hz
(d)

Frequency=18684 Hz
(e)

Fig. 4. Simulated (a) and (b) First in-Plane Modes of Vibration along *x*-Axis and *y*-Axis, respectively; (c) and (d) Non-Axisymmetric Modes and (e) Out-of Plane mode of Vibration along *z*-axis.

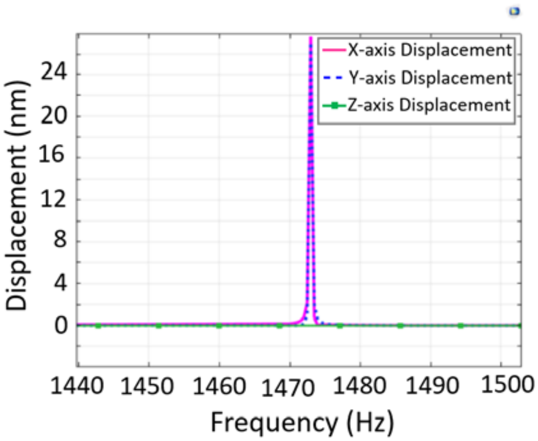


Fig. 5. x-, y- and z-Axis Proof Mass Displacement in response to a Frequency Sweep.

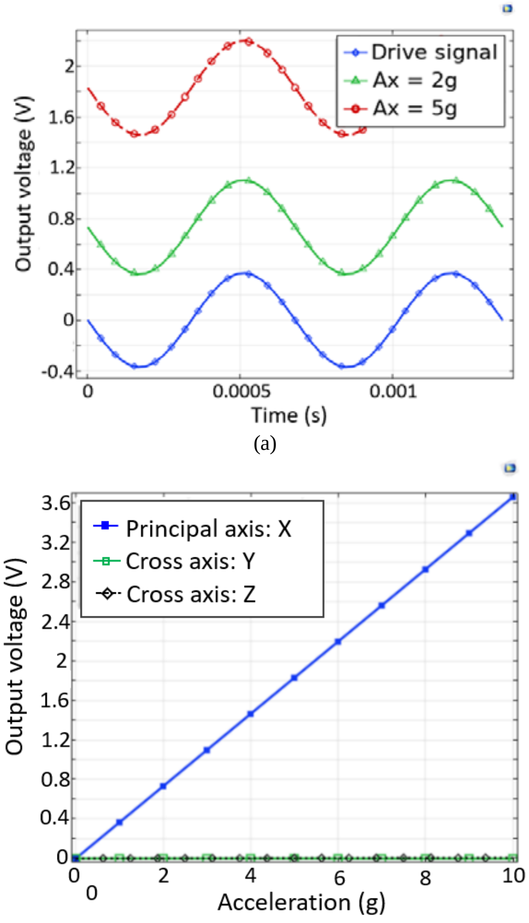
Fig. 5 shows the displacement magnitude of the lower median tip of the seismic mass as a function of a range of excitation frequency, around the resonance, using FEM simulations (under a drive amplitude voltage of 2 mV). In the simulated range, *z*-axis displacement is found with a very low magnitude that could be neglected, which confirm the proof mass oscillation in the *x/y* plane. It is worth noting that the chosen frequency of drive mode faithfully corresponds to this pic value that is near to the first axisymmetric mode of both x- and y-axis which leads to a maximum proof mass displacement while using the minimum supply voltage.

B. Mechanism of 6-DOF Detection

This section is to illustrate FEM simulation results of the studied IMU design. An investigation of signals output nature relative to linear and angular motion detection is evaluated. Also, the sensor footprint and its operating performances are extracted to be compared with other designs from literature.

C. Linear Acceleration

The analysis of voltage outputs relative to acceleration inputs is performed using a temporal FEM study [26]. The latter should be realized in practical conditions, since the structure should be permanently biased with drive voltage signals in order to be able to measure the angular velocity once applied. Upon an x-axis acceleration input, the electrical potential generated in two aligned electrodes will be opposite since one will undergo a compression and the other a dilatation. Hence, the harvested voltage is maximized by differentiating between potentials in Sx1 and Sx2 electrodes. FEM results show that the output signals, when drive mode is ongoing, give a DC offset shift proportional to the experienced acceleration superimposed to the AC signal corresponding to the drive mode reference signal (Fig. 6a). Thus, a conditioning circuitry is needed to eliminate the permanent AC signal generated by the drive mode. Processed signals offset relative to x-axis acceleration input is depicted in Fig. 6b showing a sensitivity of 364 mV/g. Similarly, for y-axis acceleration, output signals are obtained by using a differential potential between Sy1 and Sy2. Since the design represents a radial (or rotational) symmetry around z-axis, y-axis acceleration generates the same output voltages values as found with the application of x-axis acceleration inputs. For an acceleration along z-axis, the whole structure will undergo a vertical displacement, therefore the voltage generated on the four electrodes will have the same value and vary in the same direction (+ or –). Hence, output signal is collected from the sum of the four sense electrodes. Fig. 7b exhibits numerical simulations in response to z-axis acceleration input with the presentation of the resulted cross-axis signals showing a sensitivity of 65.5 mV/g.



(a)

(b)

Fig. 6. FEM Simulation Result of Output Voltage (S_{x1}-S_{x2}) in response to x-Axis Linear Acceleration Input under (a) Drive Mode and (b) Static Study.

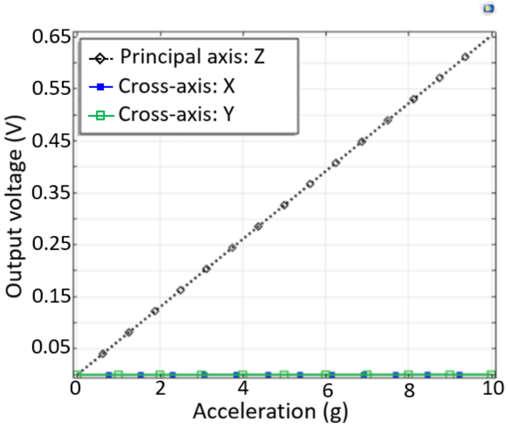


Fig. 7. FEM Simulation Result of Output Voltage in response to *y*- and *z*-Axis Linear Acceleration Input showing a Good Cross-Axis Sensitivity.

D. Linear Cross-Axis Sensitivity

Cross-axis sensitivity in accelerometers is considered as an essential property especially for high performance applications. It is defined as:

$$Cross\_Sens = \frac{\sqrt{S_{c1}^2 + S_{c2}^2}}{S} \times 100\% \qquad (11)$$

where S is the measured sensitivity in the considered direction and S_{c1} and S_{c2} are the measured sensitivities in the two other cross-axis directions.

The proposed novel design presents an excellent characteristic in term of cross-axis sensitivity; the structure gives 0.05% as cross-axis sensitivity for transversal axes and 0.02% for z-axis. Thus, it is efficiently ensured that the fact of detecting the acceleration motion in a principal axis will not affect the other axes. This crucial property is obtained thanks to the highly symmetrical structure of the model that can significantly restrain the cross-axis sensitivity without affecting the sensor sensitivity making the proposed design a highly performance inertial sensor.

E. Angular Rate

When the sensor is subjected to an angular motion, Ω, a Coriolis force, F_c, will be generated as given by:

$$\vec{F_c} = 2m\vec{v} \times \vec{\Omega} \qquad (12)$$

where m is the proof-mass mass, and v is the applied velocity component of the oscillator (displacement derivative included as well in *x/y*-plane due to the drive mode). When a longitudinal angular rate input acts (Ω_x or Ω_y), two components of Coriolis force are generated since *z*-axis is perpendicular to the *x/y* oscillation plane. So, both *x*-and *y*-axis angular velocity will be detected by the production of Coriolis force along z-axis, as:

$$F_{cz} = -2mv_y\Omega_x \qquad (13)$$

$$F_{cz} = 2mv_x\Omega_y \qquad (14)$$

Thus, in presence of a longitudinal angular rotation, the planar circular oscillation trajectory of the proof mass will be perturbed instantly due to the sinusoidal Coriolis force along z-axis. In fact, an oscillation along *z*-axis will be added, and therefore, the oscillation plane will have a well-defined inclination with respect to the initial horizontal x/y-plane (Fig 8). Hence, this inclination induces a change in the amplitude of the generated AC signal compared to that from the drive mode. In response to Ω_x, the difference in amplitudes between drive and sense signals recovered from the sense electrodes placed in y-direction (S_{y1}-S_{y2}) will be significantly higher (Fig. 9a) than the one recovered from electrodes along x-axis (S_{x1}-S_{x2}) as depicted in Fig. 9b. This means that Ω_x readout will occur only from S_y electrodes, and inversely for Ω_y. This fact leads to good cross-axis sensitivities in both x- and y-axis. In practice, total voltage is measured from the summation of potential difference between x-sense electrodes (S_{x1}-S_{x2}) and y-sense electrodes (S_{y1}-S_{y2}). Besides, for x-axis angular input, we must pay attention to the presence of y-axis component of the Coriolis force. This latter will be produced due to the presence of z-axis velocity (even with an extremely low value) that will also be combined with x-axis angular input (F_{cy}=2 m v_z Ω_x). Consequently, the seismic mass will then undergo a significant additional oscillation stretched in the y-direction that will be manifested in an elliptical curve oscillation instead of a perfect circular trajectory.

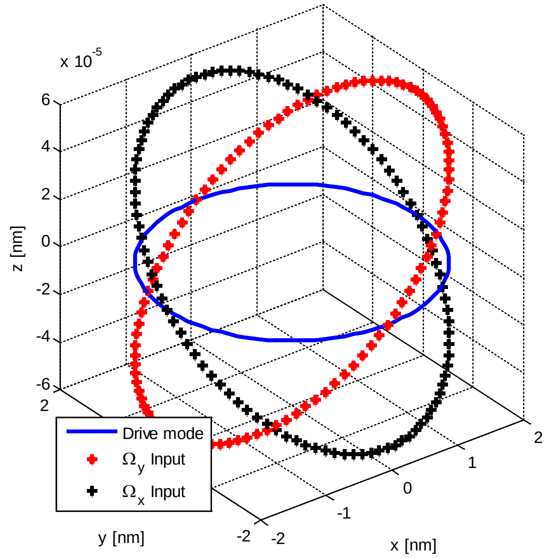
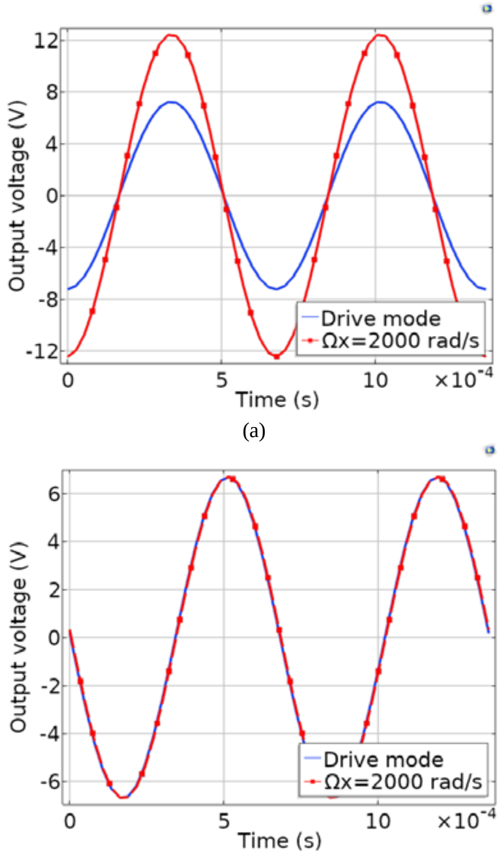


Fig. 8. Illustration of the Proof Mass Trajectory under Drive mode and in response to x- and y-axis Angular Rate Input.



(a)

(b)

Fig. 9. Output Voltages Obtained in response to an x-axis Angular Rate from (a) $S_{y1} - S_{y2}$, and (b) $S_{x1} - S_{x2}$.

When a z-axis angular rate input acts on the sensor, two components of Coriolis force are generated, and the oscillating sensor will be disturbed by a supplementary vibration in the x- or y-axis directions due to the generation of Coriolis effect exhibited by:

$$F_{cx} = 2mv_y\Omega_z \tag{15}$$

$$F_{cy} = 2mv_x\Omega_z \tag{16}$$

Using FEM simulation results, we can extract the sensor response under Ω_z input through the evaluation of the effect of Coriolis force produced in both x- and y-directions.

Consequently, the diameter of the circular oscillation trajectory of the seismic mass will either increase or decrease (Fig. 10). Therefore, the amplitude of the sinusoidal component of all sense electrodes will either increase or decrease, depending on the direction of the angular rotation.

In Fig. 11 output voltages versus time are extracted from the potential difference between Sx and Sy electrodes showing an equal amplitude value. Minimum and maximum angular rate values that the proposed sensor can linearly read in transversal and longitudinal axes are respectively ± 2000 rad/s and ± 5 rad/s. Angular rate sensitivities are numerically evaluated in all three-dimensional axes and are listed in Table I.

F. Discussion, Design Footprint and Comparison with other Designs from Literature

Detailed performances and characteristics of the novel IMU design and other piezoelectric inertial sensor designs from literature are illustrated in Table I. In term of linear acceleration detection, results reveal that the new proposed design allows a higher sensitivity compared to other accelerometer designs reported in [27, 28].

For angular rate sensing, our proposed structure gives considerable sensitivity value especially when compared with Chang et al. [29]. Besides, among the reported works, the novel IMU design exhibits the highest angular rate sensitivity along z-axis.

Main specifications of the developed IMU design are summarized in Table II as sensor footprint.

In our proposed design, for both linear and angular inputs, voltage output signals will be collected from the four sense electrodes. Changes in every electrode for each input case (linear and angular motions) are illustrated in Table III.

This illustration will help to design the circuitry of separation of the 6-DOF from the IMU and ensure the concurrent detection. It is worth noting that the drive mode will induce an initial voltage on sense electrodes with an amplitude ‘A’. ‘φ’ denotes the phase of the signal and ‘ofs’ is the offset value of the signal.

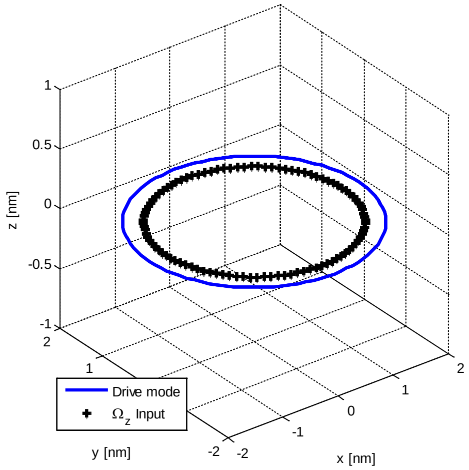
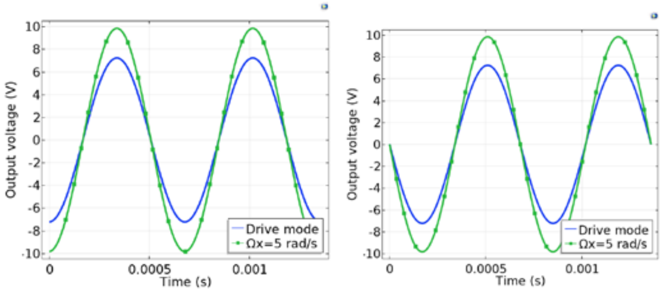


Fig. 10. Illustration of the Proof Mass Trajectory under Drive Mode and in response to *z*-axis Angular Rate Input.



(a) (b)

Fig. 11. Output Voltage Measured in response to z-axis Angular Rate Input from (a) $S_{x1} - S_{x2}$ and (b) $S_{y1} - S_{y2}$.

TABLE I. CHARACTERISTICS AND PERFORMANCES IN TERM OF ACCELERATION AND ANGULAR RATE DETECTION BY THE PROPOSED DESIGN AND OTHER PIEZOELECTRIC DESIGNS REPORTED FROM LITERATURE

Design	Dimensions scale	Piezo-electric Material	Sensitivity			
			Acceleration		Angular rate	
			A_x & A_y	A_z	Ω_x & Ω_y	Ω_z
Novel IMU	Mesoscale	PZT	364 mV/g	65.5 mV/g	2.59 mV/rad//s	522 mV/rad/s
Wang et al. [27]	MEMS	ZnO	16.3 mV/g			
Shen et al. [28]	MEMS	PLZT	31.15 mV/g			
Chang et al. [29]	Mesoscale	PZT			1.82 μV/deg/s	
Fujii et al. [30]	MEMS	PZT			25 mV/deg/s	
Parent et al. [31]	Mesoscale	PZT			3 mV/deg/s	
Okada et al. [10]	Mesoscale	PZT	300 mV/g		2.5 mV/deg/sec	
LV *et al.* [32]	Mesoscale	Quartz	0.695 pC/m/s²	0.423 pC/m/s²	0.028 pC/rad/s²	0.023 pC/rad/s²

TABLE II. NOVEL IMU DESIGN FOOTPRINT

Concurrent 6-DOF detection Linear acceleration and angular rate sensing
High sensitivity Up to 522 mV/rad/s
Very low cross-axis sensitivity Up to 0.02 %
IMU based on a single-core design instead of two sensors

TABLE III. OUTPUTS IN ALL SENSE ELECTRODES IN RESPONSE TO A POSITIVE INPUT RELATIVE TO ALL 6-DOF INPUTS

	S_{x1}			S_{x2}			S_{y1}			S_{y2}		
	Amp	φ	ofs	Amp	φ	ofs	Amp	φ	ofs	Amp	φ	ofs
	A	π	0	A	0	0	A	-π/2	0	A	π/2	0
A_x	A	π	-	A	0	+	A	-π/2	0	A	π/2	0
A_y	A	π	0	A	0	0	A	-π/2	-	A	π/2	+
A_z	A	π	+	A	0	+	A	-π/2	+	A	π/2	+
Ω_x	A	π	0	A	0	0	A↑	-π/2	0	A↑	π/2	0
Ω_y	A↑	π	0	A↑	0	0	A	-π/2	0	A	π/2	0
Ω_z	A↑	π	0	A↑	0	0	A↑	-π/2	0	A↑	π/2	0

V. 6-DOF SEPARATION CIRCUITRY

In the presence of acceleration or angular velocity components, initial voltage from drive voltage will be superimposed to the sensing signals on the same sense electrodes. The main function of the developed IMU sensor is to deal with linear and angular motion concurrently, using a single resonant structure. Hence, the challenge consists mainly in the differentiation of the linear and angular measured outputs from only four electrodes in order to be able to identify the nature and magnitude of the applied input. The used processing signal bloc diagram, operating on sensing electrodes is exhibited in Fig. 12, and signal evolution on the used sense electrodes combinations for 6-DOF inputs, are illustrated in Table IV. Basically, any accelerations will change the offset of the signals in the electrodes, while the angular accelerations will change the drive mode amplitude of the sinusoidal signals.

The separation circuitry proposed for linear acceleration detection from the four electrodes is given in Fig. 13. The principle is based on the detection of the signal offset, proportional to the acceleration input, which is the key data to differentiate between signals coming from angular or linear inputs. For *x*-and *y*-axis linear acceleration, a differential operational is used to generate S_x (=S_{x1}–S_{x2}) and S_y (=S_{y1}-S_{y2}) based on signals collected from the four electrodes. For z-axis acceleration, a summing amplifier is used to generate $S_{x1}+S_{x2}+S_{y1}+S_{y2}$. The acceleration detection circuit has the role to eliminate the AC drive signal and keep only the offset voltage. Both Signals from amplifiers are passed initially through a low pass filter with a cut-off frequency much lower than the sinusoidal drive frequency (i.e. 1471.5 Hz). The low pass filter is used to keep only the offset value of the signal. Lastly, the detected signal is passed through a voltage follower for impedance matching and for preventing loading effect [27]. The overall transfer function of the *x*-and *y*-axis acceleration detection circuit is given by:

$$H(jw) = \left(\frac{R2}{R1}\right)\left(\frac{1}{1+jR3C1\omega}\right) \tag{16}$$

The output signal will be divided to be integrated into the acceleration detection circuit and the angular rate detection circuit. The angular rate detection circuit is divided into two stages. In the first stage, shown in Fig. 13, a coupling capacitor is used to eliminate the DC offset from S_x/S_y signals. Next, a differential amplifier serves to remove the permanent AC signal, belonging to the drive mode, from the output signal provided by S_{x1}/S_{y1}. Then, a peak detection circuit is used to store the positive peak amplitude of the difference signal. The peak detector uses an operational amplifier connected in series with diode and a capacitor (Fig. 13). It is worth remembering that the initial voltage depends only on the drive input signal amplitude, so a calibration step is needed at the outset. Hence the output signal represents a DC voltage proportional to the angular rate input recorded in S_x/S_y electrode. The proposed circuit depicted in Fig. 13 is useful for differentiating an acceleration data from an angular one as illustrated in Fig. 14. To identify which angular rate input (i.e. Ω_x, Ω_y or Ω_z) represents this information, a second stage needs to be added.

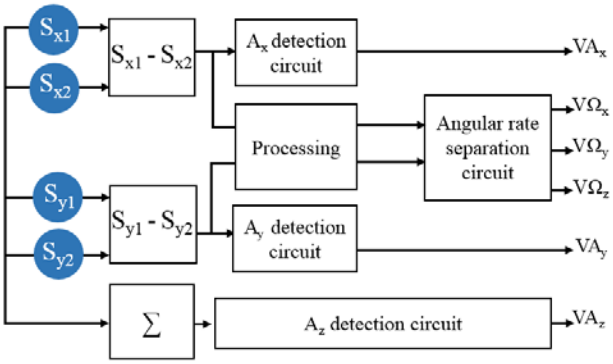


Fig. 12. Bloc Diagram of Signal Separation Circuit between the 6-DOF Inputs.

TABLE IV. OUTPUTS IN ALL SENSE ELECTRODES IN RESPONSE TO A POSITIVE INPUT RELATIVE TO ALL 6-DOF INPUTS.

	S_{x1}- S_{x2}			S_{y1}-S_{y2}			S_{x1}+ S_{x2}+ S_{x1}+ S_{x2}		
	Amp	φ	ofs	Amp	φ	offset	Amp	φ	ofs
A_x	B	π	↗	Zero signal			Zero signal		
A_y	Zero signal			B	-π/2	↗	Zero signal		
A_z	Zero signal			Zero signal			C	π	↗
Ω_x	B	π	0	B↑	- π/2	0	C ↑	π	0
Ω_y	B↑	π	0	B	- π/2	0	C ↑	π	0
Ω_z	B↑	π	0	B↑	- π/2	0	Zero signal		

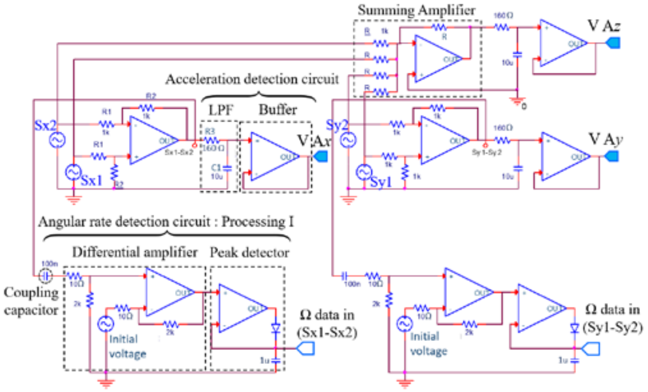


Fig. 13. Diagram of the Simultaneous Detection Circuit of Linear Acceleration and Angular Velocity Signals.

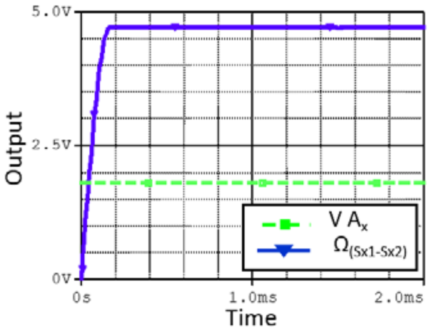


Fig. 14. Output Signals Collected from VA_x Output upon an x-axis Linear Acceleration of 5g, and from $\Omega(S_{x1}$-$S_{x2})$ Output upon an Angular Velocity of 2000 rad/s.

Outputs relative to *x*-, *y*-, and *z*-axis angular rates are recognized based on the data collected simultaneously from $\Omega_{(Sx1-Sx2)}$ and $\Omega_{(Sy1-Sy2)}$ outputs. Indeed, as depicted in Table IV, Ω_z is detected only if the same data are collected in both outputs $\Omega_{(Sx1-Sx2)}$ and $\Omega_{(Sy1-Sy2)}$. Whereas, Ω_x is detected when considering only a change in $\Omega_{(Sy1-Sy2)}$ signal and the change in only $\Omega_{(Sx1-Sx2)}$ is devoted to Ω_y detection. Thus, the angular rate dissociation circuit is designed as follow: $\Omega_{(Sx1-Sx2)}$ and $\Omega_{(Sy1-Sy2)}$ data are compared in parallel with two op-amp voltage comparators ($\Omega_{(Sx1-Sx2)}$ is on the positive input of the first comparator, and $\Omega_{(Sy1-Sy2)}$ is on the positive input of the second one). The two outputs of the comparators are then passed through a logic NOR gate. The appropriate signal output is realized by two analog demultiplexers whose outputs are controlled by the NOR gate. At the moment of same data are collected from $\Omega_{(Sx1-Sx2)}$ and $\Omega_{(Sy1-Sy2)}$ outputs, the input of each demultiplexer is transmitted to the 'Y1' pin. Then, by using a non-inverting summing amplifier followed by a buffer, the output voltage corresponding to Ω_z is faithfully gathered from the two demultiplexers. In the opposite case ($\Omega_{(Sx1-Sx2)}$ and $\Omega_{(Sy1-Sy2)}$ are different), $\Omega_{(Sx1-Sx2)}$ and $\Omega_{(Sy1-Sy2)}$ data are transmitted to 'Y0' pin of each demultiplexer, and the signal will refer to either Ω_x or Ω_y angular rate depending on which signal is non-zero as explained before. The block diagram and simulation results of the proposed circuit are respectively depicted in Fig. 15 and Fig. 16.

The used circuit for angular rate dissociation represents a very efficient method for signal separation since no leakage of angular rate readout occurs in all three axes. This fact makes the proposed 6-DOF IMU design a highly performed sensor showing practically no cross-axis in all 6-axis. Other particular interest is that under no circumstances, angular outputs disturb or affect acceleration ones and vice versa. This fact is ensured thanks to the use of the specific electronic circuitry. Further works can shed light on the study of a self-exciting drive mode circuit in order to ensure perfect oscillation conditions.

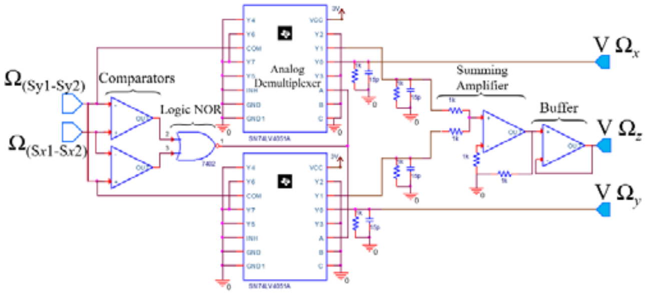


Fig. 15. Bloc Diagram of Separation Circuit between x-, y-, and z-axis Angular Rate Input.

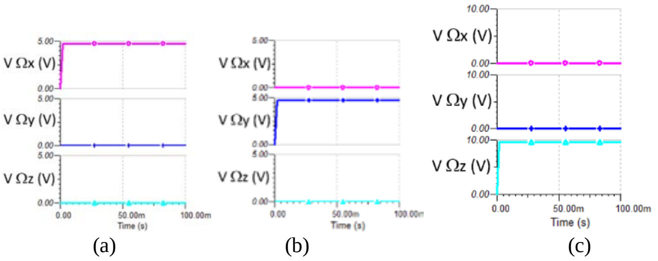


Fig. 16. Output Simulation from (a) Ω_x, (b) Ω_y, and (c) Ω_z Inputs using the Proposed Circuitry.

VI. Conclusion

This paper reports the operation principles of 6-DOF detection using a new piezoelectric IMU design based on a single proof mass oscillator. The major asset is to combine acceleration and angular rate sensing into a single-core device. The sensor performances and characteristics were investigated from finite elements simulations. Thanks to its highly symmetrical structure, the novel IMU design exhibited very low cross-axis sensitivity values and in term of sensitivity, it has shown great performances when compared with other piezoelectric designs. Linear acceleration sensitivity was found to be 364 mV/g for both *x*-and *y*-axis, and 65.5 mV/g for the *z*-axis constituent. In term of angular rate detection, the sensitivity was found to be 2.59 mV/rad/s for both *x*-and *y*-axis, and 522 mV/rad/s for the *z*-axis. A complete conditioning circuit detection was efficiently designed for the separation of linear and angular readouts using a minimal number of electrodes (only four electrodes) and the concurrent detection of 6-DOF is achieved.

References

[1] V. Saahar and R. Durai, “Designing MEMS Based Tuning Fork Gyroscope For Navigation Purpose,” International conference on Communication and Signal Processing, pp. 1102–1107, April, 2013.

[2] B. Mashadi and M. Gowdini, “Vehicle Dynamics Control by Using an Active Gyroscopic Device,” Journal of Dynamic Systems Measurement and Control, vol. 137, pp. 1–12, 2015.

[3] C. Acar, A. Schofield, A.R. Trusov, L.E Costlow and A.M. Shkel, “Environmentally robust MEMS vibratory gyroscopes for automotive applications,” IEEE Sensors Journal, vol. 9, pp. 1895–1906, 2009.

[4] B. Munoz-Barron, J. R. Rivera-Guillen, R. A. Osornio-Rios and R. J. Romero-Troncoso; “Sensor Fusion for Joint Kinematic Estimation in Serial Robots Using Encoder, Accelerometer and Gyroscope,” Journal of Intelligent & Robotic Systems, vol. 78, pp. 529–540, 2015.

[5] Z. You, “Space Microsystems and Micro/nano Satellites,” Micro and Nano Technologies Book Series, pp. 233-293, 2018.

[6] A.S. Kundu, O. Mazumder, P.K. Lenka and S. Bhaumik, “Hand Gesture Recognition Based Omnidirectional Wheelchair Control Using IMU and EMG Sensors,” Journal of Intelligent and Robotic Systems, vol. 91, pp. 529–541, 2018.

[7] E. Bostanci, B. Bostanci, N. Kanwal and A. F. Clark, “Sensor fusion of camera, GPS and IMU using fuzzy adaptive multiple motion models,” Soft Computing, vol. 22, pp. 2619–2632, 2018.

[8] D.A. Gura, G.G. Shevchenko, L.F. Kirilchik, D.V. Petrenkov and T. A. Gura, “Application of inertial measuring unit in air navigation for ALS and DAP,” Journal of Fundamental and Applied Sciences, vol.9, pp.732–741, 2017.

[9] N. Ahmad, R. A. R. Ghazilla and N. M. Khairi, “Reviews on Various Inertial Measurement Unit (IMU) Sensor Applications,” International Journal of Signal Processing Systems, vol. 1, pp. 256–262, 2013.

[10] K. Okada, T. Kakutani, H. Itano, Y. Matsu and S. Sugiyama, “Development of 6-axis Motion Sensors Using Piezoelectric Elements,” in Proceedings of the 21st Sensor Symposium on Sensors, Micromachines and Applied Systems, October 14 - 15, Kyoto, Japan, 2004.

[11] A. Parent, O.L. Traon, S. Masson and B.L. Foulgoc, “A Coriolis Vibrating Gyro Made of a Strong Piezoelectric Material,” In Proceedings of IEEE Sensors, pp. 876–879, Atlanta, GA, USA, October 2007.

[12] F. Mailly, A. Martinez, A. Giani, F. P. Delannoy and A. Boyer, “Design of a micromachined thermal accelerometer: thermal simulation and experimental results,”Microelectronics Journal,vol.34,pp.275–280,2003.

[13] S. Habibi, S. J. Cooper, J-M. Stauffer and B. Dutoit, “Gun hard inertial measurement unit based on MEMS capacitive accelerometer and rate sensor,” in IEEE/ION Position, Location and Navigation Symposium, pp. 232–237, May 2008.

[14] A. Partridge, J.K. Reynolds, B.W. Chui, E. M. Chow, A.M. Fitzgerald, L. Zhang, N.I. Maluf and T.W. Kenny, “A High-Performance Planar Piezoresistive Accelerometer,” Journal of Microelectromechanical Systems, vol. 9, pp. 58–66, March 2000.

[15] J. Wagner and J. Burgemeister, “Piezoelectric Accelerometers Theory and Application,” 6th ed., Radebeul: Manfred Weber Metra Mess- und Frequenztechnik in Radebeul e.K., 2012.

[16] A. Sharma, F. M. Zaman, B. V. Amini and F. Ayazi, “A high-Q in-plane SOI tuning fork gyroscope,” IEEE SENSORS Conference, Vienna, Austria, 24-27 Oct. 2004.

[17] Q. Li, D. Xiao, X. Zhou, Z. Hou, Y. Xu and X. Wu, "Quality Factor Improvement in the Disk Resonator Gyroscope by Optimizing the Spoke Length Distribution," Journal of Microelectromechanical Systems, vol. 27, no. 3, pp. 414-423, June 2018.

[18] D. Xiao, X. Zhou, Q. Li, Z. Hou, X. Xi, Y. Wu and X. Wu, “Design of a Disk Resonator Gyroscope With High Mechanical Sensitivity by Optimizing the Ring Thickness Distribution,” Journal of Microelectromechanical Systems, vol. 25, no. 4, pp. 606-616, Aug. 2016.

[19] Y. Cheng, W. Zhang, J. Tang, D. Sun and W. Chen, “A MEMS piezoelectric solid disk gyroscope with improved sensitivity,” Microsystem Technologies, vol. 21, pp. 1371–1377, 2015.

[20] Y. Wang, Q. Fu, Y. Zhang, W. Zhang, D. Chen, L. Yin, and X. Liu, “A Digital Closed-Loop Sense MEMS Disk Resonator Gyroscope Circuit Design Based on Integrated Analog Front-end,” Sensors (MDPI), vol. 20, 2020.

[21] J. Xie, Y. Hao, W. Yuan, “Energy Loss in a MEMS Disk Resonator Gyroscope,” Micromachines (Basel), vol. 10, Jul 2019.

[22] H. Almabrouk, S. Kaziz, B. Mezghani, F. Tounsi and Y. Bernard, “Design Presentation and Induced-Stress Study of a 6-axis Single-Mass Piezoelectric IMU,” 30th International Conference on Microelectronics (ICM), Sousse, Tunisia, 2018, pp. 144-147, 2018.

[23] H. Almabrouk, S. Kaziz, B. Mezghani, F. Tounsi and Y. Bernard, "Performance Enhancement of an Improved Design of 6-axis Single-Mass Piezoelectric IMU," 30th International Conference on Microelectronics (ICM), Sousse, Tunisia, 2018, pp. 140-143, 2018.

[24] H. Almabrouk, B. Mezghani, G. Agnus, S. Kaziz, Y. Bernard, and F. Tounsi “Geometry Investigation and Performance Optimization of a Single-Mass Piezoelectric 6-DOF IMU”, Engineering, Technology & Applied Science Research, vol. 10, pp. 6282-6289, October, 2020.

[25] F. Braghin, E. Leo and F. Resta, “The damping in MEMS inertial sensors both at high and low pressure levels,” Nonlinear Dynamics, vol. 54, pp. 79–92, Apr. 2008.

[26] H. Almabrouk, M. Hadj Said, B. Mezghani, F. Tounsi and Y. Bernard, “Simultaneous Angular and Linear Motion Detection Circuitry for a MEMS 6-DOF Piezoelectric IMU”, IEEE International Conference on Design & Test of Integrated Micro & Nano-Systems (DTS), Gammart, Tunisia, Avril, 2019.

[27] Y-H. Wang, P. Song, X. Li, C. Ru, G. Ferrari, P. Balasubramanian, M. Amabili, Y. Sun and X. Liu, “A Paper-Based Piezoelectric Accelerometer,” Micromachines (Basel), vol. 9, pp. 1–12, 2018.

[28] Z. Shen, C.Y. Tan, K. Yao, L. Zhang and Y. F. Chen, “A miniaturized wireless accelerometer with micromachined piezoelectric sensing element, ” Sensors and Actuators A, Vol. 241, pp. 113–119, 2016.

[29] C-Y. Chang and T-L. Chen, “Design, Fabrication, and Modeling of a Novel Dual-Axis Control Input PZT Gyroscope,” sensors, vol. 11, 2017.

[30] E. Fujii, R. Takayama, K. Nomura, A. Murata, T. Hirasawa, A. Tomozawa, S. Fujii, T. Kamada, and H. Torii, “Preparation of (001) Oriented Pb(Zr,Ti)O3 Thin Films and Their Piezoelectric Applications,” IEEE Transactions on ultrasonics, ferroelectrics, and frequency control, vol. 54, no. 12, Dec. 2007.

[31] A. Parent, O.L. Traon, S. Masson and B.L. Foulgoc, “A Coriolis Vibrating Gyro Made of a Strong Piezoelectric Material,” In Proceedings of IEEE Sensors, pp. 876–879, Atlanta, GA, USA, October 2007.

[32] H. Lv, L. Qin and J. Liu, “Principle Research on a Single Mass Six-Degree-of-Freedom Accelerometer With Six Groups of Piezoelectric Sensing Elements,” IEEE Sensors Journal, vol. 15, Issue 16, pp. 3301–3310, 2015.

Can Model Checking Assure, Distributed Autonomous Systems Agree? An Urban Air Mobility Case Study

Anubhav Gupta[1]
Department of Computer Science and Engineering
Florida Institute of Technology
Melbourne, Florida-32904

Siddhartha Bhattacharyya[2]
Department of Computer Science and Engineering
Florida Institute of Technology
Melbourne, Florida-32904

S. Vadivel[3]
Department of Computer Science
BITS Pilani Dubai Campus, Dubai
United Arab Emirates

***Abstract*—Advancement in artificial intelligence, internet of things and information technology have enabled the delegation of execution of autonomous services to autonomous systems for civil applications. It is envisioned, that with an increase in the demand for autonomous systems, the decision making associated in the execution of the autonomous services will be distributed, with some of the responsibility in decision making, shifted to the autonomous systems. Thus, it is of utmost importance that we assure the correctness of distributed protocols, that multiple autonomous systems will follow, as they interact with each other in providing the service. Towards this end, we discuss our proposed framework to model, analyze and assure the correctness of distributed protocols executed by autonomous systems to provide a service. We demonstrate our approach by formally modeling the behavior of autonomous systems that will be involved in providing services in the Urban Air Mobility framework that enables air taxis to transport passengers.**

***Keywords*—Formal methods; autonomous systems; distributed algorithms; assurance for distributed protocols; distributed protocol modeling and verification; distributed autonomous systems**

I. Introduction

Advancement in technologies associated with autonomous systems have significantly increased the use of autonomous systems in day to day activities. Additionally, communication capabilities have enabled the use of multiple autonomous systems to be used for executing autonomous missions. Unmanned Aerial Systems (UAS) are used across diverse applications, such as structural health monitoring [1], data driven path planning [2], and object classification [3]. Research by Cesare and Hollinger presented in [4] explores execution of multi-UAS missions under unreliable communication and limited battery life, for search and rescue applications that include urban search and rescue, military reconnaissance, and underground mine rescue operations.

With the increase in UAS applications several research efforts have started focusing on handling contingency scenarios such as investigating emergency landing for UAS by evaluating data available from population census and occupancy estimates from mobile phone activity [5]. Additionally, Automatic Supervisory Adaptive Control (ASAC) method enables the UAS to fly with a damaged wing [6]. As the applications start focusing on safety critical operations it becomes evident that we need to develop and deploy methods and frameworks for assuring multiple autonomous systems working together can complete the operations successfully.

One of the essential elements of an intelligent system design is in the formulation of the logic to intelligently respond to the environment. We in this research effort, focus on representing the logic as in artificial intelligence that enables automated reasoning to verify the correctness of the design. The automated reasoning involves the utilization of theories in formal methods, which is a branch of artificial intelligence that allows the design of logic as models on which we can execute queries, that prove through automated searches if the design satisfies the required properties.

This paper describes work on the verification and assurance of agreement among UASs by designing and implementing a distributed protocol with a case study for Urban Air Mobility (UAM) [7]. The implementation of the logic involved in distributed reasoning and its verification is done using Uppaal [8], a real time model checking tool. In order to accomplish the goal we present a mapping of requirements as identified from UAM model, that is implemented as queries in Uppaal [8].

The rest of this paper is organized as follows. Section II of this paper talks about the previous work that has been done in the area of formal methods and distributed protocols. Section III specifically discusses the framework for the formal modeling and analysis of the behavior of autonomous systems for UAM. It also discusses the expected architecture of distributed autonomous agents providing service in UAM. In Section IV, formal modeling paradigm is discussed in detail. This section elaborates upon the mathematical representation within the modeling paradigm and formal modeling tool Uppaal [8], which is used to build the formal model for the logic involved in distributed protocol for multiple autonomous systems to cooperatively provide a service. This section also states the behavioral model of autonomous systems in Uppaal [8] and the various verification properties used to verify the model. Experimental results are presented in Section V and finally the conclusion along with future work is inferred in Section VI.

II. Literature Survey

A. Formal Methods or Assurance Methods

There has been considerable previous work done in the area of formal methods for assurance [9], [10], [11]. In [9] the research discusses a method to perform run-time assurance for learning systems with an assurance architecture designed in Architecture Analysis and Design Language (AADL) and formal contracts for each of the components modeled and verified in Assume Guarantee (AGREE) annex. In [12] Davis discusses an approach to use architectural analysis to prove that the protocol designed for multi-agents satisfies the specified properties. This effort also emphasized the use of AADL and AGREE for formal assurance. For formal assurance of cooperative agents [10], discusses the development of a framework to represent cognitive architecture which is then translated into a formal environment Uppaal to verify that the autonomous agent along with interaction with the human achieves the objective. These studies emphasizes on the fact that how the use of formal methods can greatly increase our understanding of a system by revealing inconsistencies, ambiguities, and incompleteness that might otherwise go undetected.

Further, Kern and Greenstreet utilize the emergence of formal methods as an alternative approach to ensuring the quality and correctness of hardware designs [13]. Also, they emphasize the two main aspects to the application of formal methods in a design process which are modeling a formal framework that specifies the desired properties of a design. The second and more important aspect is the verification process and tools that are used to reason about the relationship between a specification and a corresponding relationship. In [14], Devillers et al. present a formal modeling and verification approach for a leader election algorithm. It describes how formal methods is used to formally model the leader election algorithm as an I/O automaton, and then it describes the verification process to prove that the implementation matches the specification. The authors emphasize the importance and use of formal methods to increase confidence in the correctness of protocols, hardware and software systems [14].

The above-cited works depict the evolution of formal methods as a formal modeling technique over the years and why it is of utmost importance to model any hardware or software specification before deploying them in a real-world environment. Formal methods have been used over the years not only for modeling designs and software but also for verification and validation of these complex designs that help in identifying subtle errors during the design process which can be later eliminated during the implementation stage.

In formal methods, model checking or theorem proving are two of the prominent methods, that are used to verify satisfaction of properties within a designed system, where model checking is automated. Model checking is a method for checking whether a model of a system meets a given specification (correctness). This is mainly associated with hardware or software systems, where we want to check liveliness requirements, as well as safety requirements. To algorithmically solve this, both the model as well as its specifications are formulated in a precise mathematical language. A model, is generally a graph such as a state machine diagram, representing the behavior of a system. The state machine diagram includes, states, transitions, condition checks and actions associated with the transitions.

The main purpose of model checking is to examine whether the evolving traces of a model, generated as an execution tree satisfies the user-given property specification. Model checking for formal verification has been used as a successfully adjunct to simulation-based verification and testing.

B. Distributed Protocols and Analysis

Phillips in [15] describes the characteristics of distributed systems and their protocols. It specifically focuses on the client-server model which is used to develop a set of requirements for a distributed system along with a description of the architecture [15]. With the advancement of networking technologies such distributed systems have significantly grown in numbers so, it has become really important to apply formal methods to the field of distributed protocols [16] to prove that the distributed systems correctly operate to achieve the required functionality.

In [17], Bhattacharyya et al. discuss the formal modeling and verification of distributed systems modeled with quasi-synchrony. It mainly provides an intuitive modeling environment that allows specification of high-level architecture and synchronization logic of quasi-synchronous systems [17]. As an example a leader selection problem is discussed where the objective had been to verify a leader is elected among a set of autonomous systems. A more elaborate explanation of verification of quasi synchronous systems is described by Miller et al. in [18] where they discuss the importance of distributing critical systems to make them redundant and fault-tolerant so that they can meet the reliability requirements. The authors specifically describe the integration and enhancement of distributed systems with innovative formal verification tools such as Satisfiability Modulo Theories (SMT) based model checkers for timed automata to provide system engineers with immediate feedback on the correctness of their designs. This work mainly focuses on the design of distributed complex systems using formal method techniques, but our approach proposes the modeling and verification of the distributed logic required for successfully executing distributed operations autonomously. Also, [18] uses examples of quasi-synchronous systems to model and verify the Pilot Flying System, the Leader Selection Case, the Active-Standby System, and the Wheel Breaking System (WBS). In a presentation [19] by Thomas Ball from Microsoft at the NUS university recently, he explains the importance of formal methods as model checking tool for distributed systems. The presentation mainly focuses on automated checking of the complex design implementation using formal methods for infinite-state systems. It also shows the importance of automatically verifying distributed systems before they can be deployed so that they are provably correct. It also talks about how formal methods find bugs in system designs that cannot be found through any other known technique.

The work in [20] exhibits a methodology to develop mathematically checkable parameterized proofs of the correctness of fault-tolerant round-based distributed algorithms. It focuses on how to replace informal and incomplete pseudo code by syntax-free formal and complete definitions of a global-state transition system. In [21], Fakhfakh et al. discuss various formal verification approaches for distributed algorithms. The study shows how there has been a rapid increase in the field of distributed algorithms due to the advances in networking technologies. It also provides information for researchers and developers to understand the contributions and challenges of the existing formal verification technologies for distributed algorithms and paves the way to enhance the reliability of these distributed algorithms [21]. In [22], the work focuses on how formal methods can be used to analyze, design, and verify security protocols over open networks and distributed systems.

As we can see that there has been considerable work done in the field of distributed protocols and formal methods [16]. But none of the work specifically focuses on modeling the logic of distributed autonomous systems using formal methods for UAM. Our **contribution** has been in the design of a framework that can be applied to the formal modeling and verification of logic designed for distributed autonomous systems to successfully execute services. We have also formally mapped the requirements for autonomous services to prove that the distributed autonomous systems have a consensus among themselves. We also propose an architectural representation of how autonomous services can be designed and verified before deployment.

III. Framework

Fig. 1 shows the process flow diagram for formal verification of distributed protocol for multiple distributed autonomous systems. The process starts with stating the requirements i.e. the goal that needs to be satisfied by the distributed autonomous systems. The requirements in our research flow from the emerging services provided by autonomous systems such as, Last Mile Delivery [23], Air Taxi and Air Metro [7]. Among these services Urban Air Mobility [7] is a futuristic concept that is being researched and developed all around the world. As a result, there is an immediate need for research thoroughly investigating possible scenarios for such emerging technologies which are agnostic to the actual implementation, but helps the process of identifying the infrastructure and correctly specifying the logic involved in the successfully deploying distributed autonomous operations. Our framework describes such an approach to design and implement behavioral models for autonomous systems, that can be formally verified and is independent of the technology to implement it. The formally verified models will help to deploy trusted, secure and reliable autonomous systems in real-world environment.

These requirements led to the generation of a Formal Model designed as automata and formal properties defined or stated in temporal logic. The formal model is developed using a formal verification tool called Uppaal [8]. Uppaal has an inbuilt simulator and verifier to simulate and verify the behavior of models (in our case autonomous system). Verifier aids the process of identifying errors in the model by executing properties that generate a counterexample along with a simulation trace if the property is violated, which helps to rectify the generated errors. This process is repeated until all the errors along are identified and corrections made. The verifier also helps to list and model many path properties that help in verifying various behavioral characteristics of the stated model, which otherwise is hard to identify and verify. Once the final verification is done and all the errors have been removed, we have a model that is formally verified and the logic of which can be trusted based on the formal verification. It is envisioned that this model can then be translated into a graphical simulation environment in order to see the exact behavior of autonomous system and also to generate real-time data. The simulation environment can be any environment that supports the integration of multiple autonomous agents, multiple drones or VTOL Planes such as X-Plane [24], AirSim [25] or Robot Operating System (ROS) [26]. This translation from formal model to simulation environment is not realized in this research.

Fig. 2 below graphically shows a hypothesized distributed architecture to support services as expected in UAM. Fig. 2(a) shows how a city can be decomposed into zones supported by multi agent environments and it's various components. Each zone is further composed of several drones that are distributed in nature, managed by a server. There is constant interaction between the drone modules and the server within a zone. Each zone interacts with other zones present within the city with the help of servers present inside each zone. There can be multiple servers based on the requirements but, for simplicity we have defined only one in the model. This constant interaction between various zones makes it a multi agent distributed environment.

Fig. 2(b) elaborates upon each zone and describes the various components and their respective functionality in detail. The Distributed Autonomous Agent Environment (DAAE, Zone) consists of various components such as buildings or nodes from where service requests are generated, drones or agents that serve requests that are generated and server. Each individual drone also comprises of it's internal server and a module that can interact with the simulation environment. The building or the nodes are responsible for generating a service request which is then passed on to the server. Along with the request, the X and Y coordinates of the building are also sent to the server, which will later be passed on to each individual drone to compute the linear distance. The server is responsible for validating the request which is then broadcasted to all the drones in the zone. Once the request is received by all the drones, they go through various checks such as verification of sensor values, battery level, authenticity of the request etc. After successful validation of the checks, all the available drones calculate their linear distance from the node that generates a request. After calculation the drones exchange their distance to the requested node, with all other drones. All drones mutually agree upon the drone that is closest to the requesting Node. This mutual agreement without the interference or involvement of any kind of central observer makes the whole model distributed and de-centralised where the decisions are taken by drones present in the distributed system. The drone module is responsible for all the communication with the server and is also responsible to carry out various checks and the linear distance calculation. After a drone has been mutually selected to serve the request, the other drones go to the start location and are available to serve any other request in the network. The described architecture of a zone is modeled in Uppaal, where each of the components are a template/process. The behavior of each component is modeled in the formal verification tool Uppaal [8] and the verification of the logic is carried out using Uppaal verifier. The tool also helps to

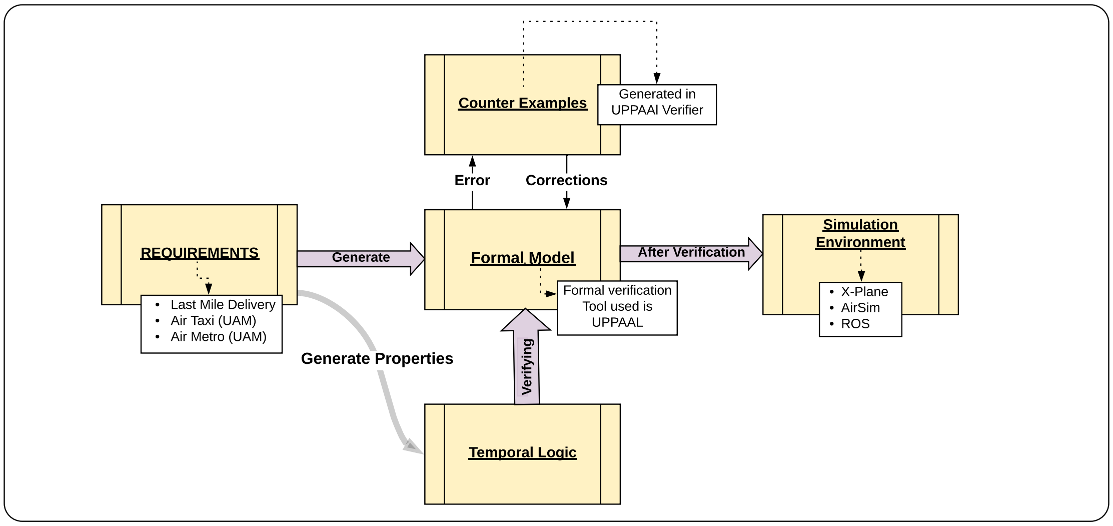


Fig. 1. Process Flow Diagram for Formal Verification of Autonomous Agents

generate simulation traces that help to identify the errors in logic which are then rectified.

This verified logic can be then translated and used to simulate this model in any real simulation environment such as AirSim [25], X-Plane [24] or ROS [26] that supports interaction of multiple autonomous agents. This real time simulation will help to generate real time data of the scenarios for a targeted service, which can be stored and later processed to improve the efficiency of the whole system. This data can also be used to develop distributed learning models for autonomous systems that will be more robust in nature and will be much more efficient. This discussed simulation is part of the future work while, this paper mainly focuses on proposing the architecture, generating formal models for the embedded components and finally, verification of the same using temporal logic to represent the requirements.

IV. Formal Modeling Paradigm

The modeling paradigm was selected after looking at several possible techniques of modeling, including Markov chains and architectural representations. We decided that the most appropriate method of representing user behavior is through the use of a Finite-State Automata (FSA) because it allows us to visualize the graphical diagram of the user's behavior easily. It enables the use of well-defined tools to perform automated analysis early in the design phase, which would empower us to reason about the logical representations of the user's behavior at the time and to evaluate alternative design options in case there were profound implications. We developed the models that are representing our knowledge base by following the principles of Finite-State automata (FSA) [27].

In order to choose the correct platform for the purpose of designing and verifying the formal model of user's behavior, several formalism such as NuSMV [28], Uppaal [8], PVS [29], and Z3 [30] were considered carefully. We chose Uppaal [8][31], due to its ability to model timing aspects that are critical for cybersecurity, as well as its ability to generate and visualize counterexamples. Uppaal represents models as timed automata, and Uppaal formalism enables compositionality supports model checking over networked timed automata using temporal logic. This modeling paradigm allows the execution of requirements as temporal logic queries to check the satisfaction of relevant safety properties exhaustively. We next describe the timed automata formalism used by Uppaal.

A. Modeling Paradigm for Timed Automata

The modeling paradigm is an extension of finite automaton with clocks, more popularly known as Timed Automata [32]. One of the tools implementing this formalism is Uppaal [8], which allows the modeling of network of Timed Automata. Clock or other relevant variable values used in guards on the transitions within the automaton. Based on the results of the guard evaluation, a transition may be enabled or disabled. Variables can be reset and implemented as invariants at a state. Modeling timed systems using a timed-automata approach is symbolic rather than explicit. It allows for the consideration of a finite subset of the infinite state space on-demand (i.e., using an equivalence relation that depends on the safety property and the timed automaton), which is referred to as the region automaton. There also exists a variety of tools to input and analyze timed automata and extensions, including the model checker Uppaal and Kronos.

- **Timed Automaton (TA)**
  A timed automaton is a tuple (L, l_0, C, A, E, I), where: L is a set of locations; $l_0 \in L$ is the initial location; C is the set of clocks; A is a set of actions, co-actions, and unobservable internal actions; $E \subseteq L \times A \times B(C) \times 2^C \times L$ is a set of edges between locations with an action, a guard and a set of clocks to be reset; and $I : L \to B(C)$ assigns invariants to locations.
  We define a clock valuation as a function $u : C \to \mathbb{R}_{\geq 0}$ from the set of clocks to the non-negative reals. Let $\mathbb{R}^C$ be the set of all clock valuations. Let $u_0(x) =$

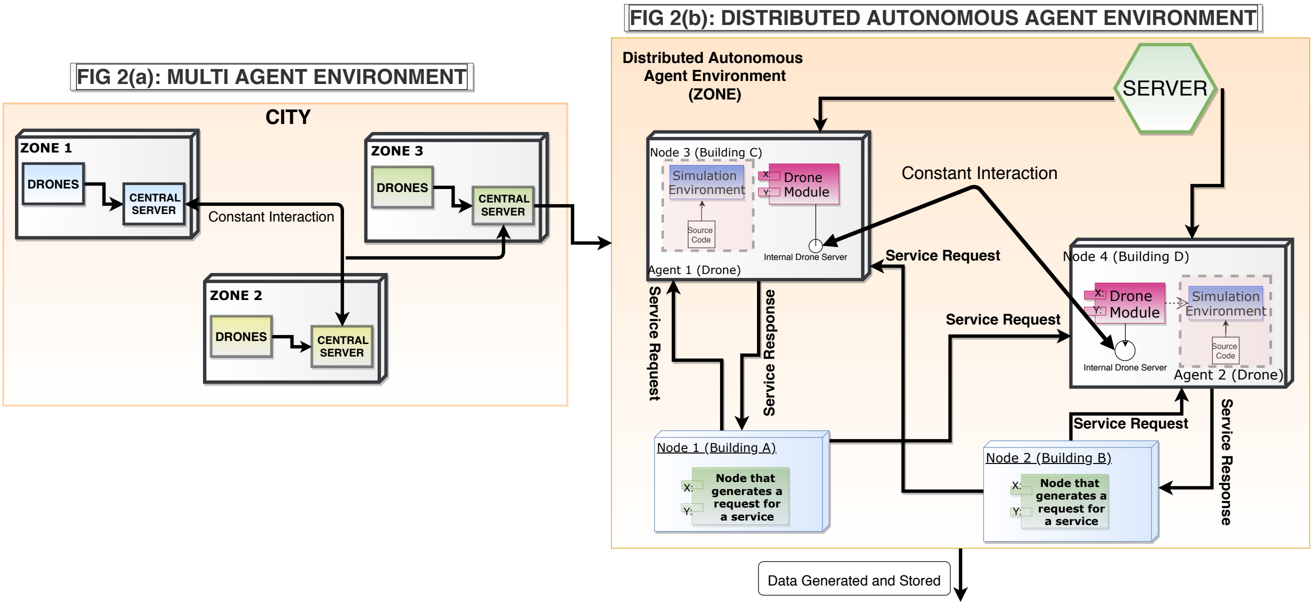


FIG 2(a): MULTI AGENT ENVIRONMENT

FIG 2(b): DISTRIBUTED AUTONOMOUS AGENT ENVIRONMENT

Fig. 2. Modeling Architecture of Distributed Autonomous Agents

0 for all $x \in C$. If we consider guards and invariants as the sets of clock valuations (with a slight relaxation of formalism), we can say $u \in I(l)$ means u satisfies $I(l)$.

- **Timed Automaton Semantics**
  Let (L, l_0, C, A, E, I) be a timed automaton TA. The semantics of the TA is defined as a labelled transition system $\langle S, s_0, \rightarrow \rangle$, where $S \subseteq L \times \mathbb{R}^C$ is the set of states, $s_0 = (l_0, u_0)$ is the initial state, and $\rightarrow \subseteq S \times \{\mathbb{R}_{\geq 0} \cup A\} \times S$ is the transition relation such that:
  1) $(l,\ u) \xrightarrow{d} (l,\ u{+}d)$ if $\forall\ d' : 0 \leq d' \leq d \Rightarrow u + d' \in I(l)$
  2) $(l,\ u) \xrightarrow{a} (l',\ u')$ if $\exists\ e = (l,\ a,\ g,\ r,\ l') \in E$ such that $u \in g$, $u = [r \mapsto 0]\ u$ and $u' \in I(l)$

  where for $d \in \mathbb{R}_{\geq 0}$, $u + d$ maps each clock x in C to the value $u(s) + d$, and $[r \mapsto 0]u$ denotes the clock valuation which maps each clock in r to 0 and agrees with u over $C \setminus r$.
  Note that a guard g of a TA is a simple condition on the clocks that enable the transition (or, edge e) from one location to another; the enabled transition is not taken unless the corresponding action a occurs. Similarly, the set of reset clocks r for the edge e specifies the clocks whose values are set to zero when the transition on edge executes. Thus, a timed automaton is a finite directed graph annotated with resets of and conditions over, non-negative real-valued clocks. Timed automata can then be composed into a network of timed automata over a common set of clocks and actions, consisting of n timed automata $TA_i = (L_i, l_{i0}, C, A, E_i, I_i)$, $1 \leq i \leq n$. This enables us to check reachability, safety, and liveness properties, which are expressed in temporal logic expressions, over this network of timed automata. An execution of the TA, denoted by $exec(TA)$ is the sequence of consecutive transitions, while the set of execution traces of the TA is denoted by $traces(TA)$.

B. *Uppaal*

Uppaal [8], an acronym based on a combination of UPPsala and AALborg universities, is an integrated tool environment for modeling, simulation and verification of real-time systems as networks of timed automata, extended with data types (bounded integers, arrays, etc.). It is used to model the logic of real time systems. For our work we have used to model the behavioral of the components for an UAM architecture [7]. We further use Uppaal [8] to verify our modeled logic in timed automata and then propose meaningful insights and results. The tool consists of three main features. First is the editor window where we model the behavioral logic for each of the modules described in detail below. Next is the simulator window, where we run a step by step simulation of the modeled logic. This helps to understand the real time functioning of the behavior of each module and further helps to refine our logic. The last and the most important part is the verifier. The verifier, utilizes a model-checker to perform an exhaustive exploration of the dynamic behavior of the system for proving safety and bounded liveliness properties. Properties are written in temporal logic to verify the logic developed. The verifier helps verify important aspects of the model and gives a deep understanding of the functioning of the model in real time scenario. It also helps to find flaws in the model that can rectified in the editor window. As a result, we are able to model a logic that has been verified and can be deployed in real time scenarios. The implemented model for UAM services as case study is described in detail in the next subsection along with detailed description and functionality of each module.

C. *Model in Uppaal*

In this subsection we elaborate on our approach to address distributed modeling and analysis for DAAE in UAM. For now, we consider that, there are three drones represented by a Drone module (Section 3) serving in a zone which is inside a city that has many such similar zones along with a Server module (Section 2), Sensor module (Section 4) and an Input module (Section 1). All these specific modules have specific roles and functions in the UAM architecture whose behavioral logic has been modeled in Uppaal. Three instances have been created for the drones in the system declaration since, all the three drones are assumed to have similar behavior for now. Algorithm 1, maps the step by step behavioral logic for drone module in Uppaal. The request is generated by a random function by the input module. The request is sent as a synchronisation event by the input module to the server module. Along with the synchronisation action, coordinates of the requesting node or building are also sent to the server module. The server module then processes the request and broadcasts it to all the individual drones available to server a request. The drone before receiving a request, checks for all sensor values using the help of sensor module (Section 4). After all checks have been performed, they then process the request received and mutually elect a drone that will serve the request without the interference of the Server module. This process of mutual selection makes the whole UAM architecture distributed and decentralised in nature. The design and functionality of each individual module along with their role in the whole behavioral model is described below:

1) **Input Module**
The instances of request are generated by the Input module. It is the one responsible for generating a random request which then goes as a synchronisation event to the server where it is processed and broadcasted to all the drones in the environment. As seen in Figure 3, the Input module makes a random transition from the $Start$ state to $Generate\_Request$ state. This transition generates a random integer less than 100 and based on the integer generated it further makes a transition to one of the buildings in the environment i.e Building A, B or C. Through this process, we have tried to depict a random request generator which sends a request for service synchronisation command to the Server module. Along with the $request\_from\_building$ synchronisation, the Input module also sends the coordinates of the building from where the request is generated. These coordinates are further sent to each individual drone by the Server module. These coordinates are used in distance calculation of each drone from the building. After generating a random request, the Input module makes a transition back to the $start$ state to generate a new request for service. This process continues and random requests are generated which are then served by the drone.

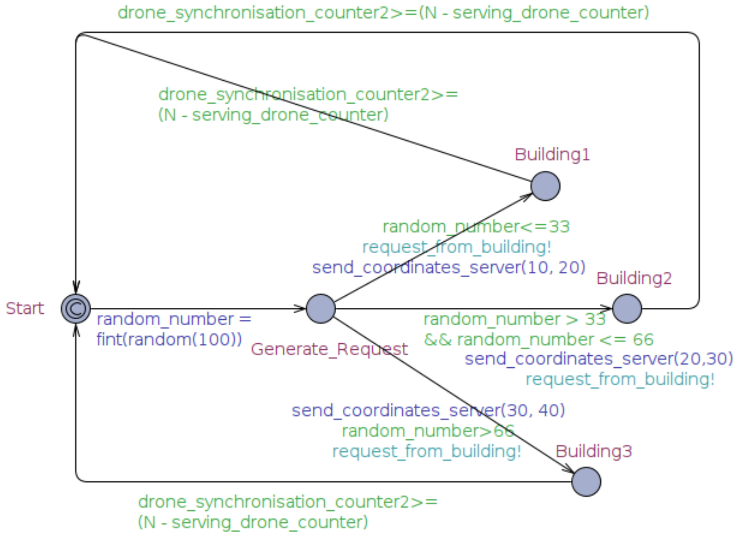

drone_synchronisation_counter2>=(N - serving_drone_counter)
drone_synchronisation_counter2>=
(N - serving_drone_counter)
Building1
random_number<=33
request_from_building!
send_coordinates_server(10, 20)
Building2
Start
random_number =
fint(random(100))
Generate_Request
random_number > 33
&& random_number <= 66
send_coordinates_server(20,30)
request_from_building!
send_coordinates_server(30, 40)
random_number>66
request_from_building!
Building3
drone_synchronisation_counter2>=
(N - serving_drone_counter)

Fig. 3. Behavioral Model in UPPAAL for Input Module

2) **Server Module**
The Server module describes the behavioral logic for Server which is responsible for routing the request generated by the Input module. As seen in Fig. 4, the Server module transitions from $Start$ state to $Wait\_For\_Request$ state when it receives a request for service from the Input module. It immediately sends a synchronisation request to the Drone module. This request goes as a synchronisation event and is received by each drone that is available to serve a request. The request is broadcasted to all the available drones along with the location coordinates of the building from where the request is generated. Only after the drones have received an authentic request from the Server module, they proceed further to calculate linear distance in-order to mutually elect the nearest drone to serve the request. At this point, the server module waits until it receives a synchronisation serve! event from the drone that is chosen to serve the request. The drone which is chosen to serve the request sends a synchronisation action to the server module indicating that, the request generated is being served by one of the drones present in the environment. Only after receiving the serve! synchronisation, it transitions from $Wait\_For\_Communication$ state to $Repeat\_Request$ state. During this transition, the time taken from the moment a request is sent and until it is accepted by the nearest drone is stored in a variable called $time\_server$. After this state, the server modules makes a transition back to the $Start$ state to process and send any other request if available to the Drone module. This process continues repeatedly until there are no more service requests.

3) **Drone Module**
The Drone Module defines the behavioral logic of drone architecture in the UAM model. There are many instances of the drone module that can be defined in the system declaration of UPPAAL editor environment. Fig. 5 below graphically shows the drone module in the UPPAAL editor window. Initially every drone is in the $Start$ state. Once the drones are ready, they go to $Ready$ state. While making the transition from $Start$ to $Ready$,certain counters are initialized. The variable i in the module represents the identification number of the drone that is being referenced. Whilst in the $Ready$ state,

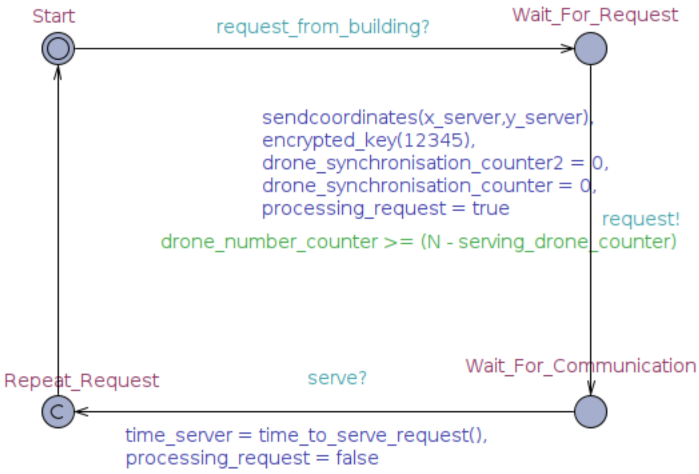

Start
request_from_building?
Wait_For_Request
sendcoordinates(x_server,y_server),
encrypted_key(12345),
drone_synchronisation_counter2 = 0,
drone_synchronisation_counter = 0,
processing_request = true
request!
drone_number_counter >= (N - serving_drone_counter)
Repeat_Request
serve?
Wait_For_Communication
C
time_server = time_to_serve_request(),
processing_request = false

Fig. 4. Behavioral Model in UPPAAL for Server Module

each drone waits for every other available drone and also waits for the Server to generate a request. Once the request has been generated, it is decrypted by each drone to check if the request is coming form authentic server or not and if proved, the drones make transition from $Ready$ to $Sensor\_Check$ state. The $Sensor\_Check$ state is where each individual drones will check if the various parameters are working normally and if the drone is in good condition to fly and serve a request. If the sensors are normal and the condition of the drone is healthy to fly it will make a transition to $Availability\_Check$ state. If any of the instrument or parameter is not working properly the drone will exit the loop by making a transition to $Report\_Error$ state. In the next state, each drone performs a linear distance calculation to calculate it's distance from the Node where request is generated. After calculating the distance, all drones will update their respective distances to a global list along with their specific identification number. After updating the distance, all drones mutually agree upon the drone which is closest to the requesting Node and select the drone nearest to the Node, to serve the request. Here the drones also perform a check for principle of quasi-synchrony [15] [17] i.e no drone should serve more than twice while others have not served once. This way all the drones get a chance to serve the requests if they are not the closest to the requesting node. This process where the drones mutually agree upon the one to serve the request without the interference of the server module or any other central module, makes the architecture distributed [17] and de-centralized in nature. After completing all these steps and mutually selecting the drone to serve a request, all drones wait at the state $Make\_Decision$ where the decision is made by each individual drone according to the mutual agreement. The drone chosen to serve the request makes a transition to the state $Serving\_Request$ while others make a transition to $Ready\_To\_Serve$. All the other drones are available again to serve any new request generated by the server module. The drone serving the request, updates certain variables and makes itself unavailable for any new request. It also sends a serve synchronisation command to the server to indicate that the request generated by the server is being served. After serving the request, the drone calculates the total time taken to serve the request and makes itself available again to serve any new request. This process continues for each and every request generated at the server side. Every time a new request is generated all the available drones perform sensor checks, authenticity check and shortest distance calculation. Always, the drone closest to the requesting node is chosen to serve the request keeping in mind the principle of quasi synchrony is satisfied [15] [17].

4) **Sensor Module**
The Sensor Module as shown in Figure 6 consists of the $Start$ and the Get_Sensor states respectively. Every time a drone transitions from $Start$ state to $Ready$ state a synchronisation event $start$ is sent to the Sensor module. The sensor module then synchronises and makes a transition from the $Start$ state to Get_Sensor state. While transitioning, it gets the latest real-time sensor values such as altitude, fuel, temperature etc. of the respective drone that sends the synchronisation and returns it to the Drone module. These values are later used by the Drone to check if it is healthy to serve the request and if all parameters are above the safe threshold limit.

D. Formal Verification Requirements

Uppaal allows for verifying requirements modeled as properties, that are useful for ensuring correctness, detecting inconsistencies, as well as flaws in the design according to the proposed modeling and analysis framework for UAM model. For example, Uppaal is capable of detecting whether there is a deadlock in the model, the results of which can further be used to find out logical flaws in the behavior of the developed model. In this subsection, we present various requirements modeled as properties, that one may want to verify with respect to UAM model, and also present meaningful insights into them along with brief description of each. The verification helps to check for any inconsistencies or flaws that may be present in the behavioral logic. After identifying the flaws, they are corrected and a consistent and a robust UAM model is presented through this work.

Requirement 1 : The existence of deadlock within the system should be verified

$$A[\ ]\ Not\ deadlock$$

This requirement is stated to check if there exists any deadlock in the system. The requirement is modeled as a property in UPPAAL verifier, that proves and presents a simulation trace of the state where a deadlock exists. After examining the particular case and scenario we find out that

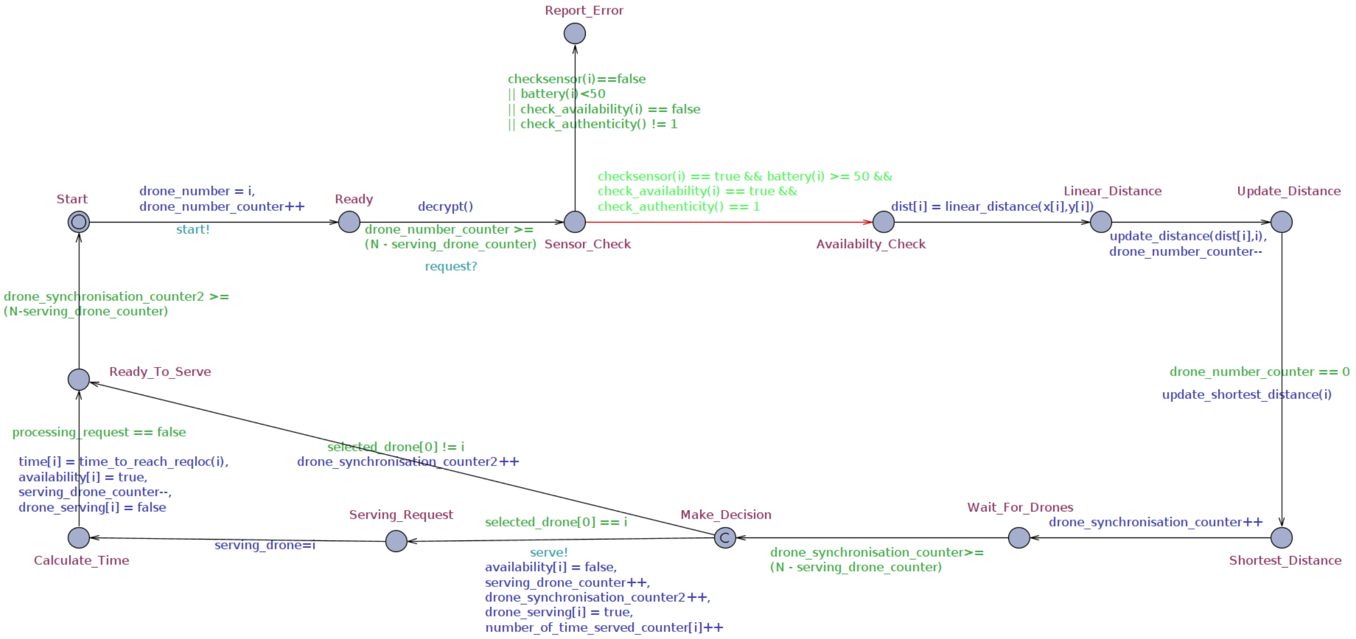

Report_Error
checksensor(i)==false
|| battery(i)<50
|| check_availability(i) == false
|| check_authenticity() != 1
Start
drone_number = i,
drone_number_counter++
start!
Ready
decrypt()
drone_number_counter >=
(N - serving_drone_counter)
request?
Sensor_Check
checksensor(i) == true && battery(i) >= 50 &&
check_availability(i) == true &&
check_authenticity() == 1
Availabilty_Check
dist[i] = linear_distance(x[i],y[i])
Linear_Distance
update_distance(dist[i],i),
drone_number_counter--
Update_Distance
drone_synchronisation_counter2 >=
(N-serving_drone_counter)
Ready_To_Serve
drone_number_counter == 0
update_shortest_distance(i)
processing_request == false
time[i] = time_to_reach_reqloc(i),
availability[i] = true,
serving_drone_counter--,
drone_serving[i] = false
selected_drone[0] != i
drone_synchronisation_counter2++
Serving_Request
selected_drone[0] == i
Make_Decision
Wait_For_Drones
drone_synchronisation_counter++
Calculate_Time
serving_drone=i
serve!
availability[i] = false,
serving_drone_counter++,
drone_synchronisation_counter2++,
drone_serving[i] = true,
number_of_time_served_counter[i]++
drone_synchronisation_counter>=
(N - serving_drone_counter)
Shortest_Distance

Fig. 5. Behavioral Model in UPPAAL for Drone Module

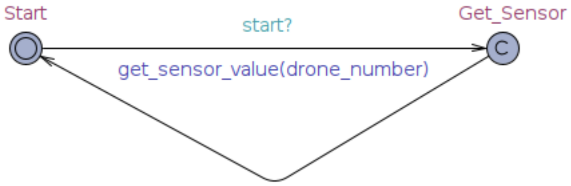


Fig. 6. Behavioral Model in UPPAAL for Sensor Module

the deadlock does not indicate any flaw or inconsistency in the logic. Similarly, for every specific system, this scenario will have to be examined to figure out if the deadlock is necessary or it is a reflection of faults and inconsistencies in the system. For example, if the service provider does not want to provide any service during night time, then a deadlock at night will indicate correct and consistent logic. Therefore, specific to the model, deadlocks have to be examined to see if it's needed or the logic has to be changed in order to remove them. **Requirement 2 : All the drones present in the system shall be able to provide service at the same time**

$E <> Drone1.Calculate\_Time$ && Drone2.Calculate_Time && Drone3.Calculate_Time

The requirement checks if all the drones present in the model can be busy at the same time to provide service to different requests i.e, all of them are servicing three individual requests simultaneously. For our model, this requirement proves indicating, that there exists a path where eventually, all three drones can be serving at the same time which shows that each drone functions independent of the other drones, but the decision are taken with mutual agreement. The following requirement also helps to justify the distributed nature of each autonomous agent in the environment that functions independent of the other agents.

Requirement 3 : An available drone shall always provide service to a generated service request

$E <> Server1.Wait\_For\_Communication$
&& (Drone1.Serving_Request or Drone2.Serving_Request or Drone3.Serving_Request)

The requirement checks if there exists a path where when a request is sent by server, it is always served by the available drones. This helps us to know that there are no neglected requests and that whenever a request is sent, it is always served and not ignored. This requirement helps to verify if any service requested is left unattended in the environment.

Requirement 4 : All the drones shall mutually agree upon who should be the service provider

$A <> (selected\_drone[0] == selected\_drone[1])$ &&
$(selected\_drone[2] == selected\_drone\ [\ 0\ ])$ &&
$(selected\_drone[0] == selected\_drone[2])$

This requirement modeled as a property proves indicating for all paths eventually, all the drones mutually agree on the drone that will provide the service. The global list $selected\_drone$ contains same elements which tells us that the service provider has been chosen with mutual consent without the interference of any external or central server. This specific requirement helps to verify that even though the proposed framework is distributed in nature but the drones take certain decisions with mutual consent without the interference of any central server or agent.

Requirement 5 : All the drones shall mutual agree upon, who is the closest to the location of reqeust

$A <> (shortest\_distance[0] == \ shortest\_distance[1]$
&& shortest_distance[1] == shortest_distance[2]
&& shortest_distance[0]== shortest_distance[2])

The above stated requirement is modeled as a property in the Uppaal verifier. We get a simulation trace indicating that for all paths eventually, all drones are able to decide upon the drone that is closest to the requesting node or building. All drones individually update the global list $shortest\_distance$ indicating the distance of the drone that is closest to the requesting building and is available to serve the request.

Requirement 6 : No drone shall provide services more than two times while other drones are idle

$E <> (number\_of\_time\_served\_counter[0] > 2 \ OR$
number_of_time_served_counter[1]>2 OR
number_of_time_served_counter[2]>2)

Using the above stated requirement, we implemented the principle of quasi synchrony. As a result, we try to check, if there exists a path where either of the drones have served more than two times, while others have not served even once. This requirement modeled as a property keeps processing and does not indicate a yes or no since it is an unbounded system. This implies it is a liveliness property and hence it does not find a state where the following condition holds true. We run the execution for almost 11,700 states till we get server connection lost error and until that time it does not hold true. In a way this implies that there isn't any path where this property holds true (i.e principle of quasi synchrony holds till the time we don't lose connection with the server) but we cannot say that for sure.

Requirement 7 : Authenticity of the incoming request from the server shall be verified by all drones

$A <> (Drone1.key == Server1.local\_key) \ \&\&$
(Drone2.key==Server1.local_key) &&
(Drone3.key==Server1.local_key)

Yes, for all paths eventually, all drones check if the request is coming from the authorised server or not. The server while sending a synchronisation service request, sends an encrypted key along with it. Each drone individually decrypts the key and compares it with the existing shared key. Only if the request is authentic, it will be served by the available drones otherwise it will be ignored.

Requirement 8 : An unavailable drone shall not process another request until it becomes available again.

$E <> ((Drone1.availability[0] == false \ \&\&$
Drone1.Make_Decision) OR (Drone2.availability[1]==false &&
Drone2.Make_Decision) OR (Drone3.availability[2]==false
&& Drone3.Make_Decision))

The above requirement tries to find if there exists a path eventually where a drone is already in the process of serving a request, goes to serve a request again i.e an unavailable drones serves a new incoming request. The requirement modeled as property keeps on running for approximately 12,065 states until connection to the server is lost. This indicates that till 12,065 states, there is no state where the above condition holds true. To prove or disprove the property we need to consider a bounded automata. Therefore, for now we cannot say for sure if the above property is true since it keeps on running in search for a simulation trace without generating a counterexample.

Property 9 : A drone with poor health shall not be chosen to serve an incoming request
The above requirement tries to find if there exists a path eventually where the battery of a drone is less than 50% and it is chosen to serve the request.

$E <> ((bat[0] < 50 \ \&\& \ Drone1.Make\_Decision)$
OR (bat[1]<50 && Drone2.Make_Decision) OR (bat[2]<50 &&
Drone3.Make_Decision))

We assume a threshold of 50% and do not want any drone with a battery value of less than the threshold to serve a request. This threshold value can be changed if needed. The property keeps on running to find a path until the server connection is lost. We need to make the model bounded in order to prove the following liveliness property. We can say that for at-least 12,458 states there doesn't exist any such path, but cannot guarantee for the whole model since it keeps on running without generating a counterexample.

Property 10 : A drone with a malfunctioned sensor should not be chosen to serve a request

$E <> ((technical\_sensor[0] == false \ \&\&$
Drone1.Make_Decision) OR (technical_sensor[1]
==false && Drone2.Make_Decision) OR
(technical_sensor[2]==false && Drone3.Make_Decision))

Through this requirement we try to investigate if there exists a path eventually, where a drone whose sensor has been malfunctioned or is not working properly, is chosen to serve the request. The requirement stated as property keeps on running until server connection is lost indicating it is unable to find such path for the number of states it runs. We need to make the model bounded to accurately indicate if it holds true or not. As of now, we cannot say for sure that the property holds true for the whole model since it keeps on running infinitely without generating a counterexample.

V. Results

This section evaluates the results of the various properties that are mentioned above. In general, we are able to verify that the distributed drones in the autonomous environment mutually agree and take decisions without the interference of any central server or module. Table 1 below evaluates the experimental results for each property. The first two columns of the Table 1 show the time taken (in seconds) by each property to execute and the total run-time memory (in megabytes) consumed. The next column indicates if the property proves or not. As we can see, some of the properties prove and some do not. Few properties keep running in loop until we get a server connection error. For these properties, we can't say for sure if they hold true or not since it is an unbounded system. The next column describes the number of states each property iterates through. Some properties that prove, iterate through all reachable states. If the verifier finds a counterexample for a particular property, it gives a simulation trace and indicates that the property does not hold true. These properties also iterate through all possible reachable states to look for a counterexample. The properties that keep on running without proving, iterate through many states as listed in the table until we get server connection error. The next column indicates if a simulation trace is generated while verifying a property. It is noteworthy that in Uppaal a simulation trace is generated when a property does not hold true i.e. the model checker finds a counterexample. Some properties which keep on running, do not generate any simulation trace and we get server connection error. An automated simulation trace is also generated when the "There exists (E<>)" property proves. Through this verification process, we are able to verify the formal behavioral logic and develop a model which is consistent and free of errors.

During the verification process, a counterexample was generated along with a simulation trace which showed that the above stated property was not satisfied. As seen in Fig. 7 the two available drones ($Drone1$ and $Drone2$) are not initialized yet since they are at $S0\_Start$ state. The Server ($Server1$) receives $request\_from\_building!$ synchronisation event from the Input Module ($Input\_Request$) indicating that a request for service has been generated, which needs to be sent to all the available drones.

The server module then broadcasts the request to all the available drones available by sending a $request!$ synchronisation. As observed, the broadcasted $request!$ synchronisation is not received by the drones since they are still at $S0\_Start$ state and hence, the service request goes unattended. The property verification process helped to identify the flaw in the logic design that the request generated by the server would sometimes go unattended and will never be served. This identification of flaw led to redesign of the logic and later we were able to rectify the logic and were able to verify the above stated property.

This specific counterexample shows how formal verification and formal model checking helps in identifying and removing flaws and inconsistencies in proposed logic during design time of complex automated systems. Fig. 7 depicts one among several counterexamples which we encountered during model checking process. The property stated during model checking intends to verify if all available drones in the system are ready to receive a request when a server is sending it.

TABLE 1. EXPERIMENTAL RESULTS OF PROPERTY VERIFICATION AND MODEL CHECKING.

Property	Time (sec)	Virtual Memory (mb)	Does it Prove?	No. of states iterated	Simulation Trace
1	0.09	10.456	No	All reachable states	Yes
2	0.017	10.452	Yes	All reachable states	No
3	0.001	10.712	Yes	All reachable states	No
4	0.001	10.712	Yes	All reachable states	No
5	0.001	10.632	Yes	All reachable states	No
6	60	1202	No, keeps running	11700	No, server connection error
7	64	1161	No, keeps running	14543	No, server connection error
8	68	1162	No, keeps running	12065	No, server connection error
9	68	1167	No, keeps running	12458	No, server connection error
10	63	1177	No, keeps running	13230	No, server connection error

A. *Discussion*

Our method provides artifacts such as models, logic, verification results as evidence that indicate satisfaction of requirements for a system, that is being designed. The evidence is obtained by, performing model checking at design time. This design time analysis also helps in clearly identifying the requirements that need to be implemented to achieve the functionality, by the creation of the model or the system, while abiding by the constraints provided by regulatory bodies.

Our approach, is agnostic to the technology that is finally used for implementation, thereby focused on identifying and representing the requirement for the problem to be solved, without getting into the complexities of implementation. It helps in generating and evaluating all the possible test cases that need to be tested for actual drones/agents to successfully execute the desired mission.

Once this is verified during design it can be implemented in any simulation environment and finally deployed on drones or autonomous agents. This method thus ensures through formal verification, the correctness of the logic designed. In our case, it verifies the logic developed in providing services by autonomous agents in a distributed environment.

B. *Limitations*

There are a few limitations with the study that has been conducted. Firstly, the Uppaal model built to represent the distributed protocol environment, needs to be evaluated for Scalability. Currently, it only represents three instances of the Drone Module. We need to evaluate and verify the behavioral logic for at-least more than ten drone modules since the Urban Air Mobility environment will consist of numerous drones. Secondly, the logic developed in the formal verification environment has not been mapped to an actual simulation environment. As part of next step of this research, we plan to map it to a simulation environment such as ROS and evaluate the performance of the distributed protocol architecture with multiple agents carrying out a specialized task within the environment. Lastly, some of the properties do not prove and keep on running, trying to find out a counterexample. This is because, currently, the model is unbounded. We need to bound the model and also come up with a more abstract representation in-order to figure out how the properties can be proved.

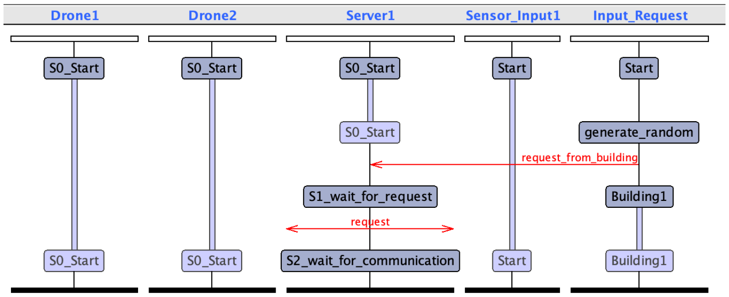


Fig. 7. Counterexample Generated during Property Verification

VI. Conclusion

Through this study, we proposed a formally verifiable framework to represent logically the behavioral that should be satisfied by the components in the infrastructure, required for distributed autonomous agents to successfully provide services. Through the property verification, we were able to prove that the distributed autonomous agents mutually agree without the interference of any central server or module. The autonomous agents are able to take decisions independently and also in synchronisation when needed. The representation is formally verified and is free of any flaws and inconsistencies.

We plan to further extend this work by incorporating scalability and heterogeneity analysis to the present study and see how heterogeneous autonomous systems behave in a distributed environment. We also plan to model and formally verify similar distributed models with other model checking tools such as, nuXmv, PRISM and see how the results vary and further try to generalise our model. Finally, we envision to map the logic from the formal model to a simulation environment.

References

[1] S. Li, Y. Wan, S. Fu, M. Liu, and H. F. Wu, "Design and implementation of a remote uav-based mobile health monitoring system," in *Nondestructive Characterization and Monitoring of Advanced Materials, Aerospace, and Civil Infrastructure 2017*, vol. 10169. International Society for Optics and Photonics, 2017, p. 101690A.

[2] C. He, Y. Wan, and J. Xie, "Spatiotemporal scenario data-driven decision for the path planning of multiple uass," in *Proceedings of the Fourth Workshop on International Science of Smart City Operations and Platforms Engineering*, 2019, pp. 7–12.

[3] Y. Qi, D. Wang, J. Xie, K. Lu, Y. Wan, and S. Fu, "Birdseyeview: Aerial view dataset for object classification and detection," in *Proceedings of IEEE GLOBECOM 2019 Workshop on Computing-Centric Drone Networks*, 2019.

[4] K. Cesare, R. Skeele, Soo-Hyun Yoo, Yawei Zhang and G. Hollinger, "Multi-uav exploration with limited communication and battery," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 2230–2235.

[5] I. O. A. Ten Harmsel and E. Atkins, "Emergency flight planning for an energy-constrained multicopter," in *Journal of Intelligent and Robotic Systems*, 2017, pp. 145–165.

[6] D. Jourdan, M. Piedmont, V. Gavrilets, D. Vos, and J. McCormick, "Enhancing UAV Survivability through Damage Tolerant Control," in *AIAA Guidance, Navigation and Control Conference*, 2010.

[7] D. P. Thipphavong, R. Apaza, B. Barmore, V. Battiste, B. Burian, Q. Dao, M. Feary, S. Go, K. H. Goodrich, J. Homola *et al.*, "Urban air mobility airspace integration concepts and considerations," in *2018 Aviation Technology, Integration, and Operations Conference*, 2018, p. 3676.

[8] J. Bengtsson, K. Larsen, F. Larsson, P. Pettersson, and W. Yi, "Uppaal: A tool suite for automatic verification of real-time systems," *Theoretical Computer Science*, pp. RS–96–58, 1996.

[9] D. Cofer, I. Amundson, R. Sattigeri, A. Passi, C. Boggs, E. Smith, L. Gilham, T. Byun, and S. Rayadurgam, "Run-time assurance for learning-enabled systems." in *NASA Formal Methods Symposium,*, 2020.

[10] S. Bhattacharyya, N. Neogi, T. Eskridge, M. Carvalho, and M. Stafford, "Formal assurance for cooperative intelligent agent," in *NASA Formal Methods Symposium LNCS*, vol. 10811, 2018.

[11] J. Rushby and R. Whitehurst, "Formal verification of al software," SRI NASA, Tech. Rep., 1989.

[12] J. Davis, D. Kingston, and L. Humphrey, "When human intuition fails: Using formal methods to find an error in the 'proof' of a multi-agent protocol." in *CAV*, 2019.

[13] C. Kern and M. R. Greenstreet, "Formal verification in hardware design: a survey," *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, vol. 4, no. 2, pp. 123–193, 1999.

[14] M. Devillers, D. Griffioen, J. Romijn, and F. Vaandrager, "Verification of a leader election protocol: Formal methods applied to ieee 1394," *Formal methods in system design*, vol. 16, no. 3, pp. 307–320, 2000.

[15] S. Phillips, "Distributed systems and their protocols," *Computer Communications*, vol. 7, no. 1, pp. 12 – 16, 1984. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0140366484900689

[16] E. M. Clarke and J. M. Wing, "Formal methods: State of the art and future directions," *ACM Computing Surveys (CSUR)*, vol. 28, no. 4, pp. 626–643, 1996.

[17] S. Bhattacharyya, S. Miller, J. Yang, S. Smolka, B. Meng, C. Sticksel, and C. Tinelli, "Verification of quasi-synchronous systems with uppaal," in *2014 IEEE/AIAA 33rd Digital Avionics Systems Conference (DASC)*, 2014, pp. 8A4–1–8A4–12.

[18] S. P. Miller, S. Bhattacharyya, C. Tinelli, S. Smolka, C. Sticksel, B. Meng, and J. Yang, "Formal verification of quasi-synchronous systems," Rockwell Collins, Tech. Rep., 2015.

[19] M. Thomas Ball, "Formal methods and tools for distributed systems," 2019. [Online]. Available: https://www.microsoft.com/en-us/research/uploads/prod/2019/01/NUS2019.pdf

[20] P. Küfner, U. Nestmann, and C. Rickmann, "Formal verification of distributed algorithms," in *IFIP International Conference on Theoretical Computer Science*. Springer, 2012, pp. 209–224.

[21] F. Fakhfakh, M. Tounsi, M. Mosbah, and A. H. Kacem, "Formal verification approaches for distributed algorithms: A systematic literature review," *Procedia Computer Science*, vol. 126, pp. 1551 – 1560,

2018, Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 22nd International Conference, KES-2018, Belgrade, Serbia. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050918314066

[22] S. Gritzalis, D. Spinellis, and P. Georgiadis, "Security protocols over open networks and distributed systems: Formal methods for their analysis, design, and verification," *Computer Communications*, vol. 22, no. 8, pp. 697–709, 1999.

[23] H. D. Yoo and S. M. Chankov, "Drone-delivery using autonomous mobility: An innovative approach to future last-mile delivery problems," in *2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 2018, pp. 1216–1220.

[24] R. Garcia and L. Barnes, "Multi-uav simulator utilizing x-plane," in *Selected papers from the 2nd International Symposium on UAVs, Reno, Nevada, USA June 8–10, 2009*. Springer, 2009, pp. 393–406.

[25] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and service robotics*. Springer, 2018, pp. 621–635.

[26] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "Ros: an open-source robot operating system," in *ICRA workshop on open source software*, vol. 3, no. 3.2. Kobe, Japan, 2009, p. 5.

[27] C. P. P., *RTL Hardware Design Using VHDL: Coding for Efficiency, Portability, and Scalability*. Hoboken, NJ: Wiley, 2006, ch. Finite State Machine: Principle and Practice, pp. 313–371, doi: https://doi.org/10.1002/0471786411.ch10.

[28] A. Cimatti, E. Clarke, E. Giunchiglia, F. Giunchiglia, M. Pistore, M. Roveri, R. Sebastiani, and A. Tacchella, *NuSMV 2: An OpenSource Tool for Symbolic Model Checking*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 359–364.

[29] S. Owre, S. Rajan, J. M. Rushby, N. Shankar, and M. Srivas, *PVS: Combining specification, proof checking, and model checking*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996, pp. 411–414.

[30] L. D. Moura and N. Bjørner, "Z3: An efficient SMT solver," in *Proceedings of the Theory and Practice of Software, 14th International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, ser. TACAS'08/ETAPS'08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 337–340.

[31] "Uppaal :A toolbox for modeling, simulation and verification of real time systems." [Online]. Available: www.uppaal.com

[32] R. Alur, C. Courcoubetis, and D. Dill, "Model-checking in dense real-time," *Inf. Comput.*, vol. 104, no. 1, p. 2–34, May 1993.

Prospects and Challenges of Learning Management Systems in Higher Education

Ahmed Al-Hunaiyyan[1], Salah Al-Sharhan[2], Rana AlHajri[3]
Computer and Information System Department Public Authority for Applied Edu. and Training (PAAET), Kuwait[1]
School of Business and IT, College of the North Atlantic – Qatar[2]
Computing Department Public Authority for Applied Edu. and Training (PAAET), Kuwait[3]

***Abstract*—Many higher education institutions nowadays are equipped with Learning Management Systems (LMS) to provide rich online learning solutions and utilize its functions and capabilities to improve the learning practices. The current study aims to gain instructors' perspective of LMS, investigate the use of its functions, and identify the barriers that may influence LMS utilization at the Gulf University for Science and Technology (GUST). This research aims to examine current practices, opinions, and challenges that help academicians and system developers contribute to better learning practices and academic achievement. The study used a quantitative method that included a sample of 58 faculty members. Findings obtained from the questionnaire indicated that instructors were generally comfortable and had positive perceptions about LMS Moodle. The results revealed that LMS's administrative functions, such as files and announcements, are widely used compared to the advanced interactive learning activities. Moreover, LMS's use on mobile devices is infrequent, and more emphasis must be placed on using LMS friendly user interfaces that can enable all tools and functions to use LMS.**

Keywords—Learning Management Systems (LMS); e-learning; Information Communication Technology (ICT); Higher Education (HE)

I. Introduction

Online learning includes many terms used in the literature, such as e-learning, distance learning, flexible learning, virtual learning, blended learning, and technology-enhanced learning [1]. E-learning involves using hardware, software, and telecommunication technologies to support and manage teaching and learning activities to transform traditional learning environments and create new and effective learning practices. The advancements of technologies helped developers provide learning management tools that encourage engaged and collaborative learning [2]. Interactive and flexible learning is a term that generally refers to activities that enhance learning opportunities and aims to create self and independent learners, while instructors facilitate the learning process [3]. Learning management systems have been employed to administer and manage online courses, track student activities, develop learning materials, deliver content to the students, monitor students' participation, and evaluate their performance [4]. Technology-enhanced learning can be merged within the classrooms, which is known as blended learning or be used to provide remote access to be part of the learning environment.

There is no single teaching method; however, some learning methods and strategies are more effective than others. Learning methods and strategies play an essential role, and significant consideration should be given to the teaching and learning style when taken from the traditional classroom and adapted to online education. Learning management systems (LMS), collaboration devices, and e-learning platforms play a vital role in allowing instructors and learners to manage, plan, deliver, and track the learning process to achieve the pedagogical objectives. Moreover, significant consideration should be given to the teaching and learning style when taken from the classroom and adapted to technical devices. Learning management systems have been implemented in many educational institutions to enhance pedagogy due to their functionalities and implementations [4]. Universities were encouraged to use LMSs to improve the collaborative environment between students and instructors. As stated by [5], most learning facilities use LMS as a tool for instruction delivery in traditional classroom settings.

LMS has been carried out since 2003 for effective teaching and learning practices. Now a day, LMSs play a vital role to achieve the learning objectives if it is used correctly; however, little attention is paid to its utilization in universities [6]. It is stated by [7] that students and instructors have the flexibility to collaborate through LMS; however, they insisted that instructors must give the support to encourage students to be actively involved in LMS. It is stressed by [8] that instructors and students rarely use the advanced functions and tools of LMS and believed that user engagement is highest for basic LMS features and lowest for features that allow interactivity, collaboration, and engagement. Since mobile devices have become ubiquitous and increasingly important, learning environments nowadays require anytime/anywhere access to course materials, collaboration, and engagement by mobile-friendly devices. Therefore, more attention needs to be paid to design friendly user interfaces for mobile devices that can encourage LMS use all tools and functions.

There are many learning management systems such as Blackboard; Moodle; Desire2Learn; Google Classroom; Schoology; TalentLMS; Canvas LMS; eCoach; A Tutor; Skillshare; LearnUpon; Edmodo. According to [9], Blackboard is the most popular LMS in the USA. Blackboard represents (33%), while Moodle (19%) is the second most popular LMS by many institutions. According to [4], LMSs may be divided into commercial and open sources. Open-source LMS can be improved and developed and then used free of charge. Examples of LMSs widely used open source is Moodle. On the other hand, commercial source LMSs are owned by private companies and are only used by registered users. The analysis of several LMSs conducted by [10] revealed that Moodle offers a wide range of features that improve pedagogical quality and includes a large number of required resources available for an online learning system. Moodle provides various functions and tools such as files, interactive lessons, folders, assignments, announcements, Hotpot quizzes, forums, chat, labels, URL links, and Turn it in [11]. These tools and functions, not only enrich managing class activities online, but also facilitate communication and collaborations between students and instructors [12].

The use of LMSs creates opportunities and helps educational institutions globally [13]. However, the adoption of LMSs requires an ongoing assessment compared to other newer technologies [4]. Various studies have been conducted to investigate the effectiveness and adoption of LMSs [4, 11, 13, 14, 15]. The current research is an attempt to gain instructors' perspective on the use of LMS, to investigate the use of its functions, to shed light on LMS opportunities, and to identify the various barriers that may influence LMS utilization at the Gulf University for Science and Technology (GUST), in Kuwait. Therefore, this research seeks to address the following questions:

1) What are the instructors' perceptions of LMS?

2) What is the instructors' use of LMS tools and functions?

3) What are the barriers to LMS utilization?

This article is structured as follows: Section 2 provides an overview of previous research, and Section 3 describes the methodology used for this study. Section 4 presents the results and discussions, while Section 5 concludes the study and proposes possible future directions.

II. Literature Review

Electronic learning is defined by [16], as an innovative learning experience that can be synchronous or asynchronous using electronic devices such as laptops, tablets, and smartphones with Internet access. Learning Management System allows instructors and students to share classroom resources, tools, and activities. According to [8] online learning is defined as a platform that facilitates the delivery and management of teaching and learning practices. LMS has tools and functions that allow schools and universities to encourage instructors to utilize them for teaching and learning processes [17], and assist them in evaluating students' activities, allowing better collaboration and interaction [18]. The proper implementation of LMS, as stated by [19], can provide students with self-paced learning, offers unlimited access to e-Learning materials, integrates social learning experiences, tracks learner progress, and increases cost-effectiveness. Furthermore, LMS offers a variety of functions and tools such as interactive books, assignments, announcements, quizzes, forums, chat, labels, and links to learning resources [19, 5, 20]. These tools and functions enrich the management of class activities online and facilitate communication and collaborations between students and faculty members [3, 21].

Many researchers believe that LMS supports teaching and learning practices. Although the study by [22] documented that LMS functions and tools help and enrich the learning environment, their research shows that LMS is primarily used as a course management tool to facilitate and enhance the learning process. Besides, [23] investigated the use of several LMS features by geology graduates at King Fahd University of Petroleum and Minerals. He used a survey to seek students' perceptions of the utilization of the LMS. The study showed that students were favoring the online discussion and believed it as a handy tool. Other tools, such as e-mail, announcement, and grade book, were also important from the students' point of view. Also, [24, 25] stressed that LMS has great potential to promote interactive, adaptive, and student-centered learning. However, the study of [24] indicated that LMS focus on the organization, management, and delivery of the learning materials. They suggest that instructors need to utilize advanced tools and activities to develop learning situations based on pedagogical approaches.

An interesting research conducted by [7] aimed to understand instructors' utilization of LMS in Malaysian HE institutions. A quantitative approach was used, in which a questionnaire was distributed to 93 instructors. The investigation included some LMS functions and tools such as announcements, files, chat, forums, exercises, and documents. The study revealed that instructors have positive perceptions of these functions; however, results showed a low percentage of instructors' utilization of LMS. Similarly, [17] examined the relationship between 222 instructors representing six Saudi Arabia universities. The investigation was conducted to understand instructors' perceptions of LMS tools and functionalities. The findings revealed that LMS capabilities were not utilized for most of the courses; however, the study indicated some barriers, such as fear of usage.

A comparative analysis of several commercial LMS was reported by [3]. The study revealed That LMSs include similar features that allow the delivery and management of different courses. However, functions that assist students and instructors in performing online laboratory experiments are not available since most engineering and science courses require these tools. The study proposed that the online learning of engineering and science courses should be facilitated via a virtual laboratory supported by LMSs. Also, [24] examined LMS usability, in which an analysis of 36 LMSs was performed. Findings have shown that all LMSs support multimedia elements such as text, files, images, audio, and videos. There is, however, a lack of communication support for LMS, which leads to using social networks outside the LMS.

Furthermore, a new study by [15] aimed at gaining student and instructor perspectives on the use of Blackboard LMS. The results showed that students were most comfortable using Blackboard and indicated that their performance and communication with instructors improved considerably. The Instructors considered the time factor to be a fundamental challenge related to the use of the LMS. Despite the challenges, however, the Blackboard Platform was a positive experience for the instructor and well received by the students.

Although LMS offers many advantages, it is argued that there is a debate about its effectiveness in education. Due to the technical requirements, LMS demands many commitments and requires technical skills from instructors [26]. Studies such as [27, 28] listed some barriers to LMS implementations: instructor commitment, lack of students, and instructors' feedback, and technical support. Research by [27] indicates that LMS platforms should be more adaptive and customized, offering easy-to-use interfaces and supporting instructors with limited technical skills. Other barriers were listed by [29] include the technical infrastructure, network capabilities, pedagogical approaches, and instructor proficiency. Concerning the LMS interactive functions [8] demonstrated that instructors and students believe that LMS enriches teaching and learning processes; however, more advanced tools and LMS interactive functions have rarely been used. As far as LMS barriers [4] administered a questionnaire to students in three universities in Saudi Arabia. The study revealed that the main obstacles to using LMSs were inadequate technical support, a negative attitude toward technology, and insufficient training on the LMS platforms. Less recognized barriers include poor Internet connectivity and networking, limited infrastructure support, lack of hardware and LMS software, and English-speaking challenges.

This article introduced an e-learning framework and implementation models to provide e-learning solutions at the Gulf University for Science and Technology. The e-learning framework was introduced by [30] to give a full-fledge e-learning implementation at GUST. The structure includes important components allowing students and instructors to access course materials through LMS anytime/anywhere. Furthermore, an e-learning implementation model was also presented by [31], which expanded the previous e-learning framework and included both the internal and external factors to implement an efficient e-learning environment at GUST. Moreover, [32] developed a new mobile learning model, which systematically integrates the mobile functions with teaching and learning practices. The model allows smooth access to the virtual learning environment, the learning management system, electronic content, and collaboration. The model also illustrates the external environment that is considered to support the model implementation.

III. Methodology

A. Survey Instrument

This study's methodological approach is a quantitative method wherein a questionnaire was designed and developed to investigate instructors' perceptions of LMS, their use of LMS tools and functions, and the barriers that may affect LMS usage by GUST instructors. The questionnaire was adapted from [7] and reviewed by experts in the field. The questionnaire is divided into four sections, section 1: demographic data, Section 2: instructors' perceptions towards LMS, Section 3: instructors' use of tools and functions provided in LMS, and Section 4: consists of questions about the barriers to the use of LMS. The items in the questionnaire consisted of a 5-point Likert type scale as 1 for strongly disagree, 2 for disagree, 3 for neutral, 4 for agree, and 5 for strongly agree, and 4-point Likert type scale as 1 for I don't use, 2 seldom, 3 for sometimes, and 4 for always. The instrument was pilot tested on a sample of 15 instructors to measure the reliability the instrument. The total score of Cronbach's Alpha is 0.901, so the questionnaire can be generalized to the primary study sample, and the results can be trusted.

B. Study Sample

This study's participants are 58 male and female instructors from the Gulf University for Science and Technology, representing around 32% of the total instructors at GUST. The authorized participants were contacted by e-mail with the information about the questionnaire. The questionnaires were sent as links by e-mail to all faculty members. Based on the questionnaires' data, statistical analysis was performed in which some analytical tools such as Frequency, Percentage, Mean, Standard Deviation (SD) were used. Data were then analyzed, as presented in the following section. The study sample distribution is illustrated in Table I, according to the demographic variables such as gender, age, and college.

TABLE I. Sample Distribution According to the Demographic Variables

Variable		**Frequency**	**Percent**
Gender	Male	40	69.0
	Female	18	31.0
Age	30 - 50 years	39	67.2
	More than 50 years	19	32.8
College	College of Business	25	43.1
	College of Art and Science	24	41.4
	Foundation program unite	9	15.5

IV. Results and Findings

This section presents the analysis results, which includes instructors' perceptions of LMS, the degree of LMS utilization among instructors, and barriers faced by instructors in the use of LMS. The mean value was used to assess instructors' perceptions.

A. Instructors' Perceptions of LMS

Table II shows the results obtained from the preliminary analysis to reflect instructors' perceptions of LMS. Instructors' responses were statistically analyzed using percentage, mean, and standard deviation (SD). Among the ten items presented in Table II, each item's mean value is higher than 3.0, which indicates that instructors have positive perceptions about LMS. Item four got the first rank with a mean value of 4.31, which demonstrates that instructors most likely want to use LMS in their teaching practices. Also, item one, "LMS helps me to communicate better with my students," got the second rank with mean value 4.22. Question 2 "The use of LMS helps me prepare learning activities" comes in third with a mean value of 4.17. The above findings are consistent with the research [15] who indicated that with Blackboard LMS, respondents' performance improved communication enhanced significantly. Given the challenges, the learning management system was a positive experience for the professors and the students. Similarly, [7] stated that students and instructors have the flexibility to collaborate through LMS, allowing better collaboration and interaction [18].

The lowest mean values ranked 9 in this section, item nine, with a mean value of 3.6, which implies that the instructors moderately believe that LMS is easy to use. The sixth element, “I like to use LMS via mobile devices,” came last, with a mean value of 3.05. This element is relatively lower than the previous items in this section, whereby the instructors do not have the confidence to use LMS with mobile devices in our case. Recent research [14] has confirmed these results. Data from the LMS transaction records were collected and carried out using log-data from different courses, showing the activities performed. The findings revealed that web LMS visits were 1,554,101 compared to 41,217 visits by mobile LMS. Remarkably, mobile, in this case, is seldom used for accessing LMS activities, which could be linked to the limited functions of mobile-based LMS. Therefore, as growing users are going mobile, LMS functions should be available and become fully compatible with mobile devices [33]. Instructors and learners can have quick and smooth access to rich services and online courses anytime and anywhere, enabling instructors to manage their learning activities effectively.

B. Instructors’ use of LMS Tools and Functions

The data presented in Table III reflect instructors’ use of LMS tools and functions. Considering the eleven items' mean values in this section, item six, “I upload files through LMS,” ranked first with a mean of 3.83. Also, item eleven, “I post announcements via LMS,” ranked second with a mean value of 3.38. Comes third item one, “I use LMS assignment function” with a mean 3.34. However, question 10, “I create digital book via LMS” and question 11, “I use chatroom with my students” ranked 10 and 11 with mean values of 1.52, and 1.40, respectively.

TABLE II. INSTRUCTORS’ PERCEPTIONS OF LMS

NO	**Item**	**Strongly agree %**	**Agree %**	**Neutral %**	**Disagree %**	**Strongly disagree%**	**Mean**	**SD**	**Rank**
1	LMS helps me to communicate better with my students	51.7	32.8	6.9	3.4	5.2	4.22	1.077	2
2	The use of LMS helps me to prepare learning activities	53.4	22.4	15.5	5.2	3.4	4.17	1.094	3
3	LMS provides effective learning for students	43.1	24.1	24.1	5.2	3.4	3.98	1.100	6
4	I will try to use LMS as part of my teaching activities	56.9	24.1	13.8	3.4	1.7	4.31	0.959	1
5	I intend to learn more about the functions and features of LMS	43.1	34.5	13.8	5.2	3.4	4.09	1.048	4
6	I like using LMS via mobile devices	19.0	13.8	32.8	22.4	12.1	3.05	1.276	10
7	LMS helps me to monitor students’ performance	36.2	32.8	20.7	6.9	3.4	3.91	1.081	8
8	LMS saves my time as an instructor	32.8	43.1	13.8	5.2	5.2	3.93	1.074	7
9	I believe that LMS is easy to use	20.7	41.4	20.7	12.1	5.2	3.60	1.107	9
10	I would recommend others to use LMS	36.2	43.1	15.5	1.7	3.4	4.07	0.953	5

TABLE III. INSTRUCTORS’ USAGE OF LMS TOOLS AND FUNCTIONS

No.	**Item**	**always**	**Sometimes**	**Seldom**	**I don’t use**	**Mean**	**SD**	**Rank**
1	I use LMS assignment functions	58.6	25.9	6.9	8.6	3.34	0.947	3
2	I use Chatroom with my students	0.0	15.5	8.6	75.9	1.40	0.748	11
3	I use Discussion/Forum function to communicate with my students	13.8	25.9	10.3	50.0	2.03	1.154	9
4	I create Quizzes via LMS	22.4	25.9	17.2	34.5	2.36	1.180	7
5	I create digital book via LMS	5.2	12.1	12.1	70.7	1.52	0.903	10
6	I upload files via LMS	87.9	8.6	1.7	1.7	3.83	0.534	1
7	I create Folders via LMS	50.0	24.1	6.9	19.0	3.05	1.161	4
8	I use Label (information) function through LMS	36.2	13.8	6.9	43.1	2.43	1.365	6
9	I post URL links through LMS	36.2	37.9	6.9	19.0	2.91	1.097	5
10	I use Turnitin to check students’ work (Plagiarism)	24.1	24.1	12.1	39.7	2.33	1.234	8
11	I post announcements via LMS	63.8	20.7	5.2	10.3	3.38	0.988	2

It is claimed by [27] that LMS includes several administrative, collaborative, and pedagogical elements that support and promote the learning process and aid in distributing online learning material. Results presented in Table III demonstrate that the utilization of LMS functions and tools varies according to the purpose of use. Administrative delivery functions are generally used, such as files, announcements, assignments, and folders, while interactive tools such as developing interactive books, chatrooms, and discussions are rarely used. This finding is consistent with [22], that LMS is widely used to focus on the delivery of learning materials rather than the creation and development of interactive practices. The study of [27, 28] documented the less use of interactive features of LMS. Having pointed this out, [6] emphasized that LMS is not only a platform for distributing learning materials but must be used appropriately to create an excellent venue for interactive and collaborative learning activities.

Previous work of [14] identified Moodle's actual use, a GUST-based learning management system, with 3600 students and 179 professors involved in this study. As shown in Fig. 1, the results showed low use of LMS interactive learning functions, such as interactive books and chat rooms, and moderate LMS administrative and management functions such as files and assignments.

C. Barriers to LMS Adoption

This section presents the results of instructors' perceptions about barriers to LMS adoption. Data were analyzed using percentage, mean, and standard deviation. The mean values, which are less than 3 in the ten items listed in Table IV, indicate that the instructors do not see the barriers as a fundamental element that hinders LMS use at GUST. Question 4 "The use of LMS via mobile devices is complicated," ranked first with a mean 2.88. The second item, which investigates students' active participation in LMS tools, got the second rank with a mean of 2.69. Comes third question 3 "LMS Interfaces confuses me" with a mean 2.53. Also, Question 10 "I did not get proper training about LMS" ranked 4, with a mean value of 2.47.

As for LMS usage via mobile devices, instructors are not sure to use LMS with mobile devices, which is consistent with a recent study by [14]. Regarding students being active in using LMS tools and functions depend heavily on instructors' attitudes toward technology. As reported by [34], some instructors do not consider LMSs' effective tools in teaching; instead, they utilize traditional strategies. For example, instead of encouraging the students to use LMS to enhance their knowledge, they undermine or avoid supporting them. Considering providing training programs, as in item 10 and the university's encouragement on the use of LMS as in item 9, revealed that instructors are not very much satisfied with training and management support. The motivation was identified as a critical factor in developing and sustaining community sense and success and achievement in an online learning environment [35].

Furthermore, it is stressed by [36, 4] that instructors should be prepared to teach, deliver, collaborate online digitized and provide learning resources, and evaluate students online. Besides, [4] concluded that preparing instructors for online teaching is a real challenge. Similarly, [37] claimed that the significant challenges of LMSs in Saudi Arabian institutions are a lack of or insufficient training and support, and infrastructure weakness in the institutions, and a lack of proper technical support.

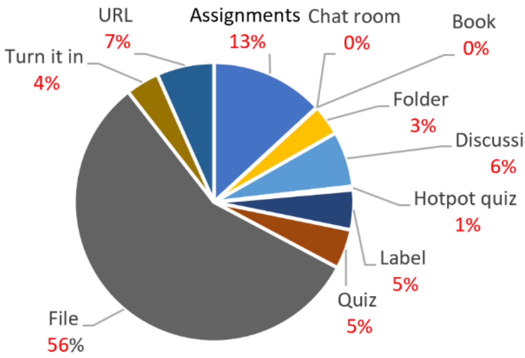


Fig. 1. LMS Functions Created by GUST Instructors [14].

TABLE IV. BARRIERS OF LMS ADOPTION

No	Item	Strongly agree	Agree	Neutral	Disagree	Strongly disagree	Mean	SD	Rank
1	I do not have enough experience to use LMS	3.4	5.2	17.2	43.1	31.0	2.07	1.006	9
2	Students are not active in using LMS tools	3.4	19.0	31.0	36.2	10.3	2.69	1.012	2
3	LMS Interfaces confuses me	3.4	13.8	31.0	36.2	15.5	2.53	1.030	3
4	The use of LMS via mobile devices is complicated	3.4	20.7	44.8	22.4	8.6	2.88	0.957	1
5	Limited Internet access restraint me from using LMS	1.7	20.7	12.1	36.2	29.3	2.29	1.155	6
6	Limited computer facilities restraint me from using LMS	1.7	13.8	10.3	44.8	29.3	2.14	1.050	8
7	Limited technical support restraint me from using LMS	5.2	6.9	15.5	50.0	22.4	2.22	1.044	7
8	There is a lack of security and privacy in using LMS	3.4	6.9	32.8	37.9	19.0	2.38	0.988	5
9	The University did not encourage me to use LMS	0.0	1.7	13.8	34.5	50.0	1.67	0.781	10
10	I did not get proper training about LMS	1.7	12.1	34.5	34.5	17.2	2.47	0.977	4

Other factors influencing the use of LMS resources, which was considered a barrier to implementation, as displayed in items 5 and 6 of Table IV, was the level of university support for internet access and the availability of computer facilities. Most instructors reported moderate and reasonable levels of internet access and computer facilities. The research findings also indicated that the university's level of technical support, security, and privacy, as shown in items 7 and 8 are not a barrier to adopting LMSs. As for other LMS implementation barriers, the study of [27] listed some challenges LMS implementations such as instructor commitment and lack of students and instructors' feedback. Moreover, [38] listed some technological barriers: infrastructure, network maintenance, and low bandwidth, which will impact the learning process with restricted access. Fear of usage was also reported as a barrier to LMS usage [17].

V. Conclusion and Future Directions

Many educational institutions use learning management system platforms are widely used by many educational institutions for administering and manage online courses creating a collaborative learning environment. Institutions should identify their needs and understand LMSs' functionality before investing in them. Barriers overcoming the implementation of LMS should be studied to ensure that all institutions will adopt a system to improve the learning process and academic performance. This research aims to examine current practices, opinions, and challenges that help academicians and system developers contribute to better learning practices and academic achievement. The study aims to gain instructors' perspective of LMS, investigate the use of its functions, and identify the barriers that may influence LMS utilization at the Gulf University for Science and Technology (GUST). A quantitative method that included a sample of 58 faculty members from GUST was used. Findings obtained from the questionnaire indicated that instructors were generally comfortable and had positive perceptions about Moodle, a GUST learning management system. The results revealed that LMS's administrative functions, such as files and announcements, are widely used compared to the advanced interactive learning activities such as interactive books and chatting. Besides, LMS's use on mobile devices is infrequent, and more emphasis must be placed on using LMS friendly user interfaces that can enable all tools and functions to use LMS. Besides, the proper adoption of LMS requires substantial technical training and encouragement by management.

Universities should encourage LMS use and focus on learning strategies through their rich tools and functions to achieve pedagogical objectives. LMS itself is not the optimal solution for student engagement in teaching and learning [39], stressing that instructors play an important part in inspiring learners to take advantage of LMS features. LMS demands a great deal of responsibility and requires instructors' technical skill due to its technical nature [38, 26]. Accordingly, universities should provide students and professors with appropriate training and guidance to use LMS tools and functions.

This study identified some minor barriers that can affect e-learning adoption, including student interaction with the system, complexity of the LMS interface, instructors' and students' readiness. There are also personal, technological, and institutional limitations. Personal includes confidence and awareness of LMS's potential and the functionality of tools and resources that enrich the teaching and learning process. Technological has the infrastructure, technical support, network bandwidth, and communication and collaboration tools to impact teaching and learning. Institutional barriers include strategic planning, management support, encouragement, motivation, and training programs to use, deliver, and develop e-learning courses.

Moreover, potential issues are cultural and social that play an essential role in accepting and adopting LMS. Instructors who have been resistant to using technology in teaching will be more likely to receive their newer teaching practices over time. Having pointed this out, educational institutions should help to teach staff to encourage the use of online instruction to be successful in their courses.

References

[1] K. H. M. Albasayna , Factors Influencing the Use of E-Learning in Schools in Crises Areas: Syrian Teachers' Perspectives, Tallinn University of Technology, Estonia, 2016.

[2] J. Zhang, D. Burgos and S. Dawson, "Advancing open, flexible and distance learning through learning analytics,," Distance Education, 40:3, DOI: 10.1080/01587919.2019.1656151, pp. 303-308, 2019.

[3] A. Aldiab, H. Chowdhury, A. Kootsookos, F. Alam and H. Allhibi, "Utilization of Learning Management Systems (LMSs) in higher education system: A case review for Saudi Arabia,"," Energy Procedia, vol. 160, pp. 731-737, 2019.

[4] A. Alenezi, "Barriers to Participation in Learning Management Systems in Saudi Arabian Universities," Education Research International. ID: 9085914. Hindawi. Accessed https://doi.org/10.1155/2018/9085914, 2018.

[5] N. Sayfouri, "Evaluation of the learning management system using students' perceptions," Medical Journal of the Islamic Republic of Iran, vol. 30, 2016.

[6] C. Chung, "Web-based Learning Management System Considerations for Higher Education.," Learning and Performance Quarterly, 1(4), p. 24–37, 2013.

[7] M. Azlim, K. Husain, B. Hussin and M. Zulisman, "Utilization Of Learning Management System In Higher Education Institution In Enhancing Teaching and Learning Process," Journal of Human Capital Developmen. Vol. 7 No. 1 January - June 2014, 2014.

[8] E. Dahlstrom, D. Brooks and J. Bichsel, "The Current Ecosystem of Learning Management Systems in Higher Education: Student, Faculty, and IT Perspectives," ECAR, September 2014. Research report: Available from http://www.educause.edu/ecar, Louisville, CO, 2014.

[9] Edutechnica, "4th Annual LMS Data Update," Edutechnica: Retrieved Dec. 2017 from: http://edutechnica.com/2016/10/03/4th-annual-lms-data-update/, 2016.

[10] C. Cigdem and G. Tirkes, "Cigdem, Cansu & Tirkes, Guzin. (2010). Open Source Learning Management Systems in Distance Learning," Turkish Online Journal of Educational Technology. volume 9 Issue 2, pp. 175-184, 2010.

[11] A. Lopes, "Learning Management Systems in Higher Education," in Proceedings of EDULEARN14 Conference 7th-9th July 2014, Barcelona, Spain, 2014.

[12] T. Scott, "8 Important LMS Features for Your E-Learning Program," 4 January 2017. [Online]. Available: https://technologyadvice.com/blog/human-resources/8-important-lms-features/.

[13] Y. Kats, , Learning Management System Technologies and Software Solutions for Online Teaching: Tools and Applications, Hershey, PA, USA: Information Science Reference, 2010.

[14] S. Al-Sharhan, A. Al-Hunaiyyan, R. Alhajri and N. Al-Huwail, "Utilization of Learning Management System (LMS) Among Instructors and Students," in Advances in Electronics Engineering, Lecture Notes in Electrical Engineering, vol 619, Singapore, Springer, 2020.

[15] J. Uziak, T. Oladiran, E. Lorencowicz and K. Becker, "2018. Students' and Instructor's Perspective on the use of Blackboard Platform for Delivering an Engineering Course.," The Electronic Journal of e-Learning, 16(1), pp., pp. 1-15, 2018.

[16] V. Singh and A. Thurman, "How Many Ways Can We Define Online Learning? A Systematic Literature Review of Definitions of Online Learning (1988-2018)," American Journal of Distance Education. Volume 33, Issue 4, pp. 289-306. https://doi.org/10.1080/08923647.2019.1663082, 2019.

[17] S. Alghamdi and A. Bayaga, "Use and attitude towards Learning Management Systems (LMS) in Saudi Arabian universities," Eurasia Journal of Mathematics, Science & Technology Education, 2016, 12(9), pp. 2309-2330, 2016.

[18] N. Emelyanova and E. Voronina, "Introducing a Learning Management System at a Russian University : Students' and Teachers‘ Perceptions," The International Review of Research in Open and Distance Learning, 15(1), p. 272–289., 2014.

[19] C. Pappas, "The Top 8 Benefits Of Using Learning Management Systems," 7 January 2016. [Online]. Available: https://elearningindustry.com/top-8-benefits-of-using-learning-management-systems. [Accessed 7 January 2020].

[20] L. Bacow, W. Bowen, K. Guthrie, K. Lack and M. Long, "Barriers to Adoption of Online Learning Systems in US Higher Education," Ithaka S+ R. Retrived 10 Dec. 2017 from: http://www.sr.ithaka.org/publications/barriers-to-adoption-of-online-learning-systems-in-u-s-higher-education/, New York, 2012.

[21] P. Venter, M. Rensburg and A. Davis, "Drivers of Learning Management System Use in a South African Open and Distance Learning Institution," Australasian Journal of Educational Technology. 28(2), pp. 183-198, 2012.

[22] M. Christie and R. Jurado, "Barriers to Innovation in Online Pedagogy," European Journal of Engineering Education, 34 (3), pp. 273-279, 2009.

[23] M. Hariri , "Students' Perceptions of the Utilization of Learning Management System (LMS) Features: A Case Study of a Geology Course at KFUPM, Saudi Arabia," International Journal of Technology Diffusion (IJTD), 5(4), 2014.

[24] R. Kraleva, M. Sabani and V. Kralev, "An Analysis of Some Learning Management Systems," International Journal on Advanced Science, Engineering and Information Technology, Vol.9 (2019) No. 4 ISSN: 2088-5334, 2019.

[25] N. Manochehr, "The Influence of Learning Styles on Learners in E-Learning Environments: An Empirical Study," Computers in Higher Education Economics Review (CHEER), V 18, pp. 10-14, 2008.

[26] E. Kanninen, Learning Style and E-Learning, Tampere University of Technology, 2008.

[27] A. Almarashdeh, N. Sahari, N. Zin and M. Alsmadi, "The success of Learning Management System among distance learners in Malaysian universities," Journal of Theoretical and Applied Information Technology, pp. 80-91, 2010.

[28] N. Adzharuddin and L. Ling, "Learning Management System (LMS) among University Students: Does It Work?," International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 3, No. 3, June 2013, 2013.

[29] D. Surry, D. Ensminger and M. Haab, "A model for integrating instructional technology into higher education," British Journal of Educational Technology, vol. 36, no. 2, p. 327–329, 2005.

[30] S. Al-Sharhan, A. Al-Hunaiyyan and H. Al-Sharrah, "New Efficient Blended E-Learning Model and Framework for K12 and Higher Education: Design and Implementation Success Factors," in Proceedings of the IEEE Fifth International Conference on Digital Information Management. (ICDIM 2010), July 05-08, 2010, Thunder Bay, Canada, 2010.

[31] S. Al-Sharhan and A. Al-Hunaiyyan, "Towards an Effective Integrated E-Learning System: Implementation, Quality Assurance and Competency Models," in Proceedings of The Seventh International Conference on Digital Information Management (ICDIM 2012) 22-24 August 2012, Macau, 2012.

[32] A. Al-Hunaiyyan, A. Al-Sharhan and R. Alhajri, "A New Mobile Learning Model in the Context of the Smart Classrooms Environment: A Holistic Approach," International Journal of Interactive Mobile Technologies (iJIM). Vol.11_No.3(2017), pp. 39-56, 2017.

[33] A. Kumar, "Make Your LMS Mobile Compatible in 4 Easy Ways," 2017. [Online]. Available: https://blog.commlabindia.com/elearning-design/4-ways-to-make-lms-mobile-compatible.

[34] L. Kyei-Blankson, E. Ntuli and H. Donnelly, ""Establishing the importance of interaction and presence to student learning in online environments," World Journal of Educational Research, vol. 3, no. 1, p. 48–65, 2016.

[35] M. Hartnett, "The Importance of Motivation in Online Learning," in Motivation in Online Education, Singapore, Springer, 2016.

[36] CoSN, "COVID-19 Response: Preparing to Take School Online," 1 March 2020. [Online]. Available: https://www.cosn.org/sites/default/files/COVID-19%20Member%20Exclusive_0.pdf. [Accessed 15 April 2020].

[37] L. Smith and A. Abouammoh, Higher Education in Saudi Arabia: Achievements, Challenges, and Opportunities, Dordrecht, Netherlands: Springer, 2013.

[38] M. AlKharang, Factors that Influence the Adoption of e-Learning An Empirical Study in Kuwait. Phd. Thesis, London: Brunel University London, 2014.

[39] J. Daniels, M. Jacobsen, S. Varnhagen and S. Friesen, "Barriers to Systemetic, Effective, and Sustainable Technology Use in High School Classroom," Canadian Journal of Learning and Technology, 39(4), 2013.

Deep Learning based Approach for Bone Diagnosis Classification in Ultrasonic Computed Tomographic Images

Marwa Fradi[1], Mouna Afif[2], Mohsen Machhout[3]
Monastir University, Physic Department of Faculty of sciences of Monastir, Tunisia

***Abstract*—Artificial intelligence (AI) in the area of medical imaging has shown a developed technology to have automatically the true diagnosis especially in ultrasonic imaging area. At this light, two types of neural networks algorithms have been developed to automatically classify the Ultrasonic Computed Tomographic (USCT) images into three categories, such as healthy, fractured and osteoporosis bone USCT images. In this work, at first step, a Convolutional Neural Network including two types of CNN models such (Inception-V3 and MobileNet) are proposed as a classifier system. At second step, an evolutionary neural network is proposed with the AmeobaNet model for USCT image classification. Results achieve 100% for train accuracy and 96%, 91.7% and 87.5% using Amoebanet, Inception-V3 and MobileNet respectively for the test accuracy. Results outperforms the state of the art and prove the robustness of the proposed classifier system with a short time process by its implementation on GPU.**

***Keywords*—USCT; Inception-V3; MobileNet; Ameobanet-V2; classification; accuracy; transfer deep learning**

I. Introduction

Nowadays, medical image classification presents the key technique for Computer Aide Diagnosis using deep learning approaches such as CNN, ANN and Evolutionary Neural works. In previous times, several types of research had been done to automatically classify medical images to have the true class diagnosis with the same way as a specialist would, especially in the area of ultrasound images classification [1]. Recently, artificial intelligence algorithms based computer Aided Diagnosis (CAD) have known a big revolution. Many researches have tried to create the similarity between human brain and computer machine by employing neural network algorithms as well as deep learning techniques. Nowadays, neural networks present an important tool for extracting information from medical images by classifying them automatically with a short time process. Moreover, Deep Learning has made a very huge development in medical image analysis and then in ultrasonic medical image classification to identify automatically a complete piece of information [2].

II. State of the Art

Given the difficulty to classify medical images automatically researchers have carried on the process of image classification based on deep learning programs as well as transfer deep learning [3]. In Fact, Transfer learning is the concept of training a pre-trained neural network model on our small dataset. Related works have used machine learning in the beginning, then transfer deep learning for medical image classification, segmentation and object detection moving from a pre-trained model, trained on big data to a small data. In recent works [4], a classifier based on fractional Fourier transform combined with SVM has been proposed to classify IRM brain images into pathological and healthy brain images. However, performance results are good, It was shown that the proposed architecture was adapted only for a small number of dataset [4, 5]. However, CNN is used in the classification of numbers and recognition of numbers with Leucin [6]. The approach of transfer deep learning was based on employing a pre-trained network, which was trained on a large number of samples for a similar task, for a new task with little annotated image data. In 2012, in [7]. Krizhevsky has published the first deep learning model with an error rate of 15.3 however Google Net has achieved 6.7% error rate in 2014. Indeed various CNN optimized algorithms have been used such as LeNet, AlexNet, ZF Net, Google Net, VGGNet and ResNet in medical image analysis, proving their efficiency. Therefore deep learning algorithms especially the convolutional neural network are rapidly emerging as an efficient method for medical image classification and then the fast diagnosis detection [8]. MobileNet has been used in [9] for skin lesion classification giving promising results with high accuracy, specificity and F-1 score, thus these results were improved by the big data augmentation and up-sampling process. Moreover, in [10] authors have used collected X-rays images, to be classified for covid-19 detection, using deep learning algorithms, thus they achieved the best accuracy, sensitivity, and specificity obtained is 96.78%, 98.66%, and 96.46% respectively. Also, MobileNet has been used in [9] for skin lesion classification giving promising results with high accuracy, specificity and F-1 score, thus these results were improved by the big data augmentation and up-sampling process. In [11], authors have employed a deep model and statistic feature fusion for feature extraction with a multilayer perceptron for medical image classification giving high classification results. However, in [12], a transfer learning has been implemented on two convolutional neural network models such as VGG16 and InceptionV3, for Pneumonia Detection. Accordingly, deep neural network-based methods [13] provides high performance in classifying the images according to the extracted features solving the issue of handcrafted feature extraction. In other hand, in previous works methods based on artificial neural networks (ANNs) have attracted more attention especially, in the area of ultrasonic images classification [14]. The Artifical Neural networks shows its huge role in medical image diagnosis class detection. Accordingly, it has been implemented in [15] for breast cancer class detection, showing a high accuracy results which proves the interest of the ANN use. To summarize and taking in consideration the achieved results in the state of the art, CNN models have proved to get good results in terms of classification accuracies and can be implemented on embedded systems like MobileNet, in contrast to the Evolutionary neural networks, which gives excellent results higher than the results achieved by a CNN, but it cannot be as a real time application given its huge size and the complexity of its architecture.

Our approach in this paper consists of a data augmentation of the number of USCT images to achieve a big data augmentation by pre-processing algorithms. Thus we solve the issue of USCT data unvailibility [16]. Then, a transfer deep learning models have been done such as Convolutional Neural Networks models (Inception V3, MobileNet-v2 and Evolutionnary Neural Network model such as AmoebaNet are applied on our dataset. Thus, the aim of this work is automatically classify USCT images with the same way a clinician would, and enables them to get the true diagnosis in short time process.

III. Method

The proposed system classifier for USCT images class diagnosis detection is depicted and detailed in Fig. 1.

A. USCT Pre-Processing

USCT image pre-processing is with huge interest given the ambiguity of the obtained USCT images, similarly to computed tomography (CT) images, where noise removal shows to be a primordial step to begin with [17,18] using Fuzzy approaches. This process leads to get a computed tomography image augmentation.

Big data augmentation is an approach to overcome the challenges posed by a limited amount of annotated training data. Augmentation is performed artificially generating more annotated training data, typically by mirroring and rotating the original images given that the difficulty to have a big number of ultrasonic computed tomographic images, image pre-processing to augment the number of images was a crucial step that should be done. For this fact, we have augmented the image number from 30 images to 250 images with a size of 256*256 by morphologic algorithms such as dilatation and erosion, then by simple rotation and finally by image processing algorithms such as Haar Wavelet transform, which is in more details in [19], K-Means and Ostu method.

1) K-Means algorithm: It is based on grouping similar data points into clusters. There is no prediction involved. Its algorithm is illustrated by these steps [20].

- Fix the k number of cluster values.
- Identify the k cluster centers.
- Determine the cluster center.
- Determine the pixel distance for each cluster center.
- If the distance is close to the center value, budge to that cluster.
- Otherwise, move to the next cluster and Re-identify the center.

The pre-processing step has an interesting effect on USCT dataset augmentation as depicted in Fig. 2.

2) Haar wavelet: A low-pass filter application remains to get an L image which is compressed and the application of a high-pass filter leads to obtain an H image which introduces image details. This process is as depicted in Fig. 3. It is done by equations (1) and (2) as follows:

$$YH[k] = \sum_n x[n]\, G[2k-n] \quad (1)$$

$$Yl[k] = \sum_n x[n] H[2k-n] \quad (2)$$

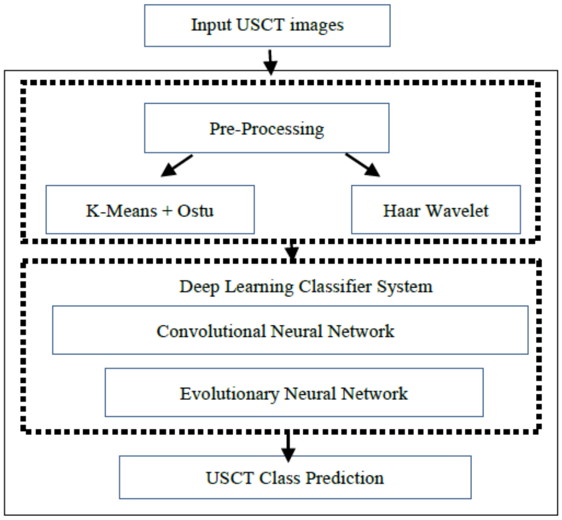

Input USCT images
Pre-Processing
K-Means + Ostu
Haar Wavelet
Deep Learning Classifier System
Convolutional Neural Network
Evolutionary Neural Network
USCT Class Prediction

Fig 1. Synoptic Flow of the Proposed Method.

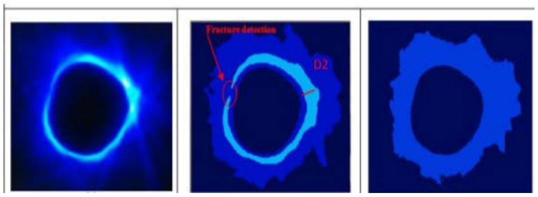


Fig 2. Augmented USCT Dataset by k-Means Algorithm.

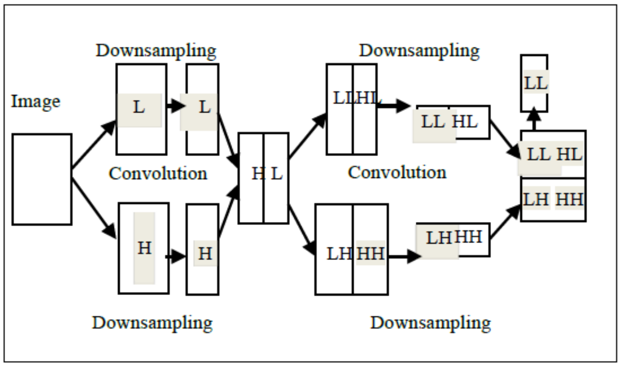


Fig 3. Principle of Haar Wavelet Application.

3) Ostu-method: The Otsu method of the threshold is the most powerful and global threshold method. It performs image binarization based on the histogram image shape. It assumes that the image for binarization contains only foreground and background pixels Using the simple formula in the Ostu algorithm, we get:

$$\sigma^2 = \Psi A \, (u\,A - u)^2 + \Psi B \, (u\,B - u)^2 \qquad (3)$$

A.4. USCT augmented data

The proposed pre-processing, applied on USCT images, shows a big data augmentation and offer a free USCT database for USCT researchers given the challenge of USCT dataset as shown in Fig. 4.

B. *Deep Learning Classifier System*

1) Convolutional Neural Network

a) Data Training

- Frameworks: We have used 250 images divided into three classes, a Linux operating system, NVIDIA Graphic Interface named Titan. With GPU computing deep learning is 10 to 30 times faster than on CPU.

- Librairies: For the data training we have needed many librairies that should be installed such as TensorFlow, keras, Numpy, Matplotlib and OpenCv.

- Training Parameters: We use the TensorFlow library and train the networks with stochastic gradient descent, with learning rate 10−3, momentum 0.9, no weight decay and batch size of 10images per step. There is no need for jittering as instead of data augmentation we can simply generate more synthetic training data. Input images are resized to 256 × 256. Deep Learning Neural Networks Models have been applied with a modification in the last layer by introducing our fully connected neural network layer. We have three classes USCT images with a size of 256*256, the first was healthy images, the second osteoporosis images and the last one was the fractured images.

b) Transfer Learning: The transfer learning is essentially based on the use of pre-trained NetWork Model to try to work around the perceived requirement of a large dataset. The training parameters are in details in Table I.

The training process of our implemented deep learning algorithm is as depicted in the Fig. 5.

2) Classical Convolutional Neural Network

a) Convolutional Layer: It is a based step in Deep learning, by the application of a filter kernel with its weights value to the input image. It has two principal functions. At first step, a kernel multiplication has been done to each pixel value covered by this kernel then all image pixel values have been summed to have the feature maps then the output of the first map will be the input of the next map, repeating the same process but with a big numbers of interconnected neurons.

b) Max Pooling Layer: It is an operation based on downsampling the output convolutional images. It outputs the max value in a local neighbourhood of each feature pas it is illustrated by Fig. 6 [6].

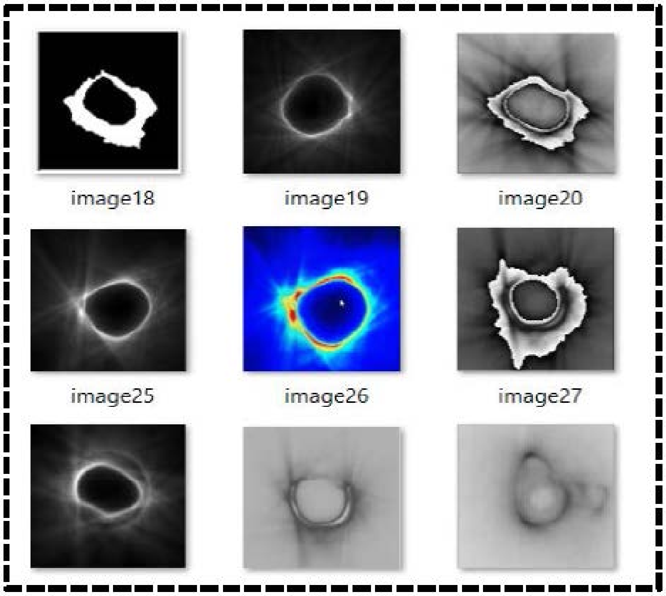
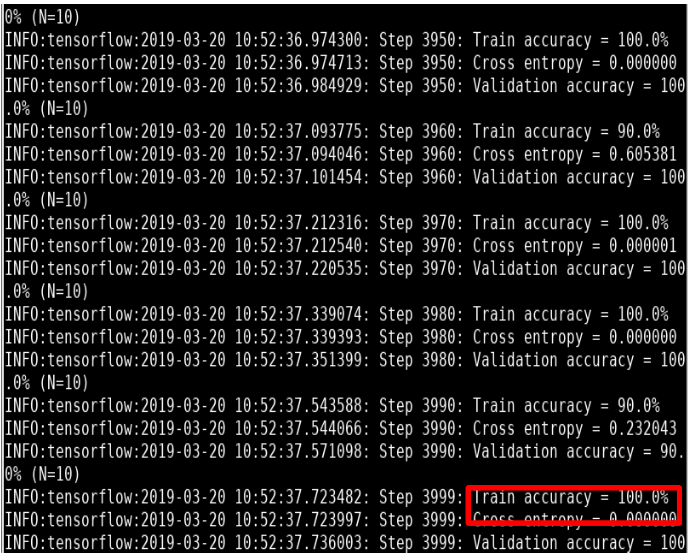


Fig 4. Augmented USCT Dataset.

TABLE I. TRAINING DATA PARAMETERS

Training Parameters	**Number**
Iterations	4000
Hidden Layers	10
Train-images	190
Validation-images	30
Test -images	30
Error-rate	0.001
Batch	10 Images/Step



```
0% (N=10)
INFO:tensorflow:2019-03-20 10:52:36.974300: Step 3950: Train accuracy = 100.0%
INFO:tensorflow:2019-03-20 10:52:36.974713: Step 3950: Cross entropy = 0.000000
INFO:tensorflow:2019-03-20 10:52:36.984929: Step 3950: Validation accuracy = 100
.0% (N=10)
INFO:tensorflow:2019-03-20 10:52:37.093775: Step 3960: Train accuracy = 90.0%
INFO:tensorflow:2019-03-20 10:52:37.094046: Step 3960: Cross entropy = 0.605381
INFO:tensorflow:2019-03-20 10:52:37.101454: Step 3960: Validation accuracy = 100
.0% (N=10)
INFO:tensorflow:2019-03-20 10:52:37.212316: Step 3970: Train accuracy = 100.0%
INFO:tensorflow:2019-03-20 10:52:37.212540: Step 3970: Cross entropy = 0.000001
INFO:tensorflow:2019-03-20 10:52:37.220535: Step 3970: Validation accuracy = 100
.0% (N=10)
INFO:tensorflow:2019-03-20 10:52:37.339074: Step 3980: Train accuracy = 100.0%
INFO:tensorflow:2019-03-20 10:52:37.339393: Step 3980: Cross entropy = 0.000000
INFO:tensorflow:2019-03-20 10:52:37.351399: Step 3980: Validation accuracy = 100
.0% (N=10)
INFO:tensorflow:2019-03-20 10:52:37.543588: Step 3990: Train accuracy = 90.0%
INFO:tensorflow:2019-03-20 10:52:37.544066: Step 3990: Cross entropy = 0.232043
INFO:tensorflow:2019-03-20 10:52:37.571098: Step 3990: Validation accuracy = 90.
0% (N=10)
INFO:tensorflow:2019-03-20 10:52:37.723482: Step 3999: Train accuracy = 100.0%
INFO:tensorflow:2019-03-20 10:52:37.723997: Step 3999: Cross entropy = 0.000000
INFO:tensorflow:2019-03-20 10:52:37.736003: Step 3999: Validation accuracy = 100
```

Fig 5. Train Data Accuracy.

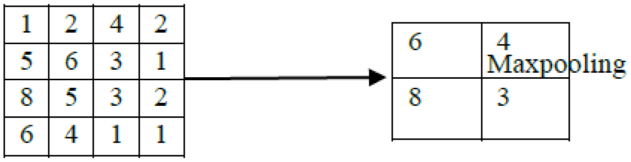


Fig 6. Max-Pooling.

c) Classifier Layer: The classifier layer is based on the fully connected layers. It plays a crucial role to classify images based the detected features. A fully connected layer is a layer whose neurons have full connections to all activation in the previous layer.

The classical CNN, consisting of a convolutional layer, a max pooling layer and a classifier layer is a depicted in Fig. 7.

3) Proposed architecture: Our proposed architecture was based on a transfer deep learning models such as AmoebaNet [21, 22], Inception-V3[1], Mobile net [23] and NasNet to classify USCT images automatically into three classes. Our approach is to modify the last fully connected layer by our FCNN layer which is composed by three categories: healthy images, osteoporosis images and fractured image. Then with our proposed architecture we have implemented it into Graphic interface processor GPU. All models have roughly the same proposed architecture by a modification of the last layer with our FCNN layer. In the example below in figure Inception –V3 architecture is detailed.

4) Transfer deep learning inception-v3 model architecture: In this part of the work, we try to transfer knowledge in order to develop a new USCT image classification system using the third generation of the inception family: “inception v3” [1].

This architecture provides 42 layers, it demonstrates more computational efficiency compared to the pervious inception family architectures. Inception v3 architecture presents a promising network composition with different parameters optimization as depicted in Fig. 8.

In order to ensure more robustness of the neural network, a 5 x 5 convolution is replaced by two 3 x 3 convolutions. By using this technique, the parameters number is decreased from 25 to 18 parameter. This technique reduces considerably the network complexity. Also, this technique contributes for the inception v3 powerful block named “inception module-A”. The inception module-A architecture is provided in Fig. 9.

One convolution of 3 x 3 is replaced by two convolutions of 1 x 3. This modification participates in the “inception module B”. In the following figure the inception module B is presented.

Another module named “inception module C is proposed in inception v3 model as presented in Fig. 11. All these modules come on reducing the number of parameters of the whole network and minimize the risk of overfitting.

5) USCT image classification using MobileNet v2: Deep convolutional neural networks (DCNN) have revolutionized computer vision area in the last few years. In this section, we aim to develop a new powerful USCT image classification and recognition system based on the lightweight neural network mobileNet v2 architecture.

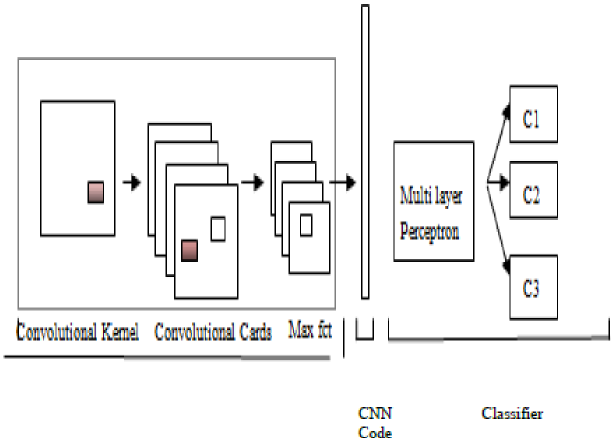


Fig 7. Classical CNN Architecture.

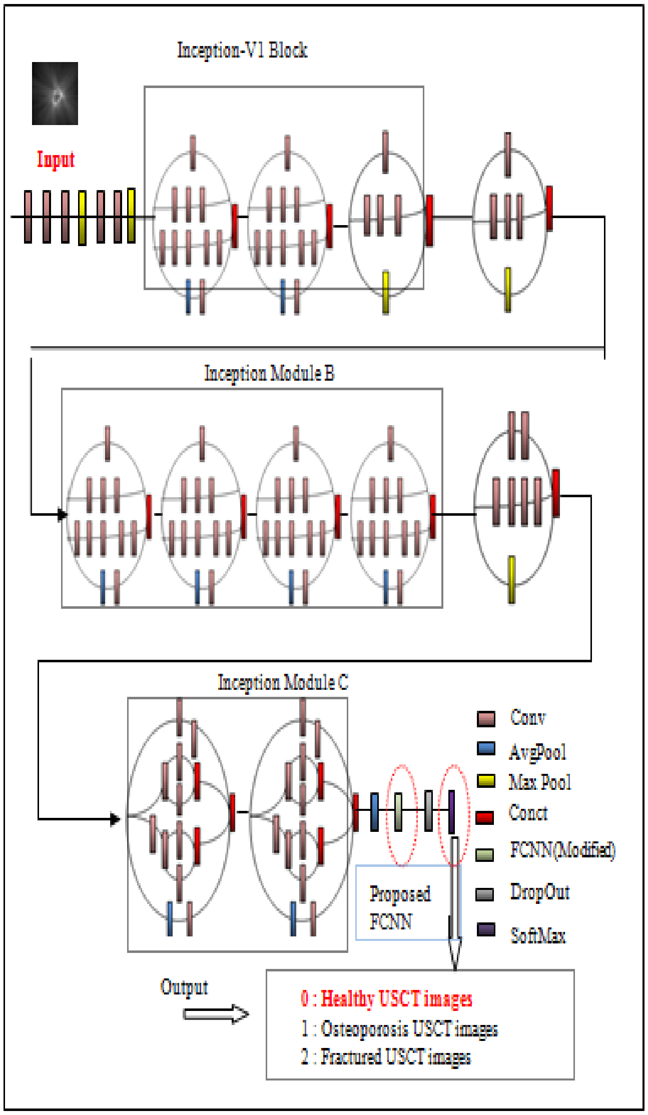


Fig 8. Proposed Inception -V3 Architecture.

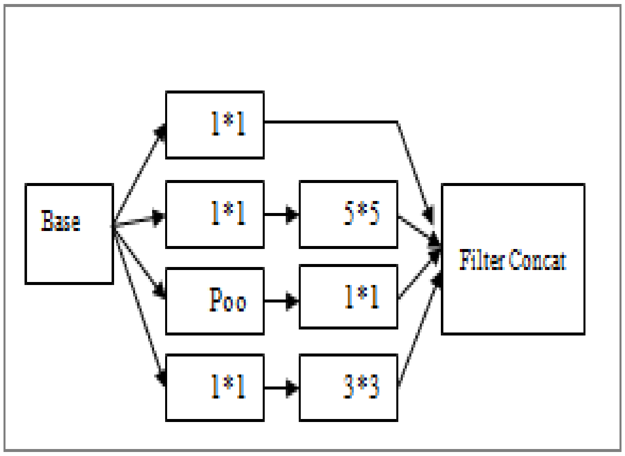


Fig 9. Inception-Module A.

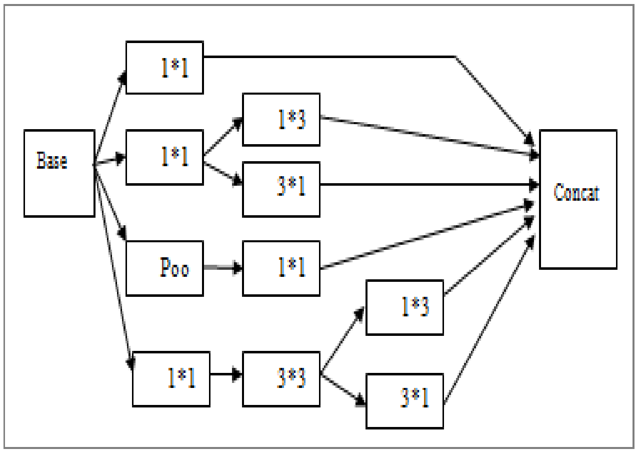


Fig 10. Inception Module B.

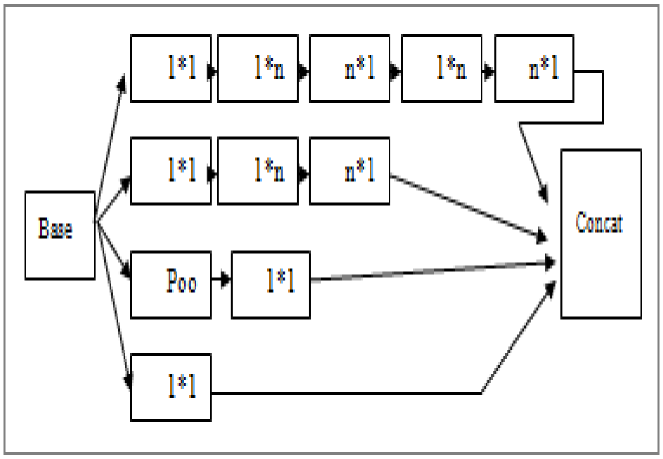


Fig 11. Inception Module C.

MobileNet v2 architecture as presented in Table II introduces a new powerful block named the "inverted residual block with linear bottleneck". All the features extracted are filtered by using a lightweight depthwise convolution. By using the depthwise convolution, we ensure a considerable reduce of the network parameters.

TABLE II. MOBILENET V2 ARCHITECTURE

Input	Layer	n
224x224x3	Conv 2D	1
112x112x32	Bottleneck	1
112x112x16	Bottleneck	2
56x56x24	Bottleneck	2
28x28x32	Bottleneck	2
14x14x64	Bottleneck	1
14x14x96	Bottleneck	2
7x7x160	Bottleneck	1
7x7x320	Conv 2D 1x1	1
7x7x1280	Average pooling 7x7	-
1x1x1280	Conv 2D 1x1	-

Another powerful fact present in the mobileNet v2 architecture is the use of the pointwise convolution before the use of depthwise separable convolution. This operation is named "bottleneck". MobileNet v2 [24] architecture present the updated version of MobileNet v1 version [23]. This updated version is promising as it provides three main new components which are the following:

- Linear bottleneck layer
- Shortcut connection
- Inverted residual block

The inverted residual layer enables the feature map to e encoded in low-dimensional sub-space. The bottleneck layer appears very similar as the residual block where each block contains the input representation followed by the explanation layer. As bottleneck layers contains the necessary information and acts as an implementation of non-linear transformation, for this fact it introduces the shortcut connection directly between bottleneck layers.

To summarize, the MobileNet v2 architecture is composed of a regular convolution followed by 11 bottleneck layers and a pointwise convolution and an average pooling layer, another pointwise convolution layer then it ends with a fully connected layer and a softmax layer used to classify the objects categories.

C. *USCT Image Classification: Approach Adopted based on Evolutionary Algorithms*

In order to improve deep learning models, an effort was devoted to apply new techniques named "evolutionary algorithms" to neural network topologies. State-of-the-art results obtained by this type of algorithms have demonstrated their higher performances compared to human-crafted ones. In this part of the proposed work, we propose to develop new USCT image classification systems using aging evolutionary algorithm Amoebanet-A [21, 22].

The AmoebaNet architecture make two additional variables to the standard evolutionary algorithms: firstly, it proposes a new tourmenant selection [25] process for evolutionary algorithms. In the standard tourmenant selection, the best neural network architecture or genotype are kept while on the AmoebaNet architecture each generated architecture or genotype is associated with a specific age parameter and bias. In this stage, the tourmenant selection is charged to keep the younger genotypes. This is the evolutionary algorithm with aging architecture technique. Secondly, AmoebaNet perform a set of mutations in a simpler way in the NasNet Search Space (NAS) [26]. NAS search space presents a space of image classifiers. It uses reinforcement learning technique to search for the best neural topologies. Applying the neural architecture search NAS to a huge dataset require a huge computational resource. To address this problem, it processes on smaller datasets and transfer the learned model on largest datasets. All architectures obtained using the NAS search space are independent and various in terms of input image size as well as the network depth.

Every neural network present in the NAS search space as shown in Fig. 12 is composed of convolution layer or (calculation cell). We note that the NAS search space search for the best calculation cell structure. By searching for the best calculation layer, the NAS search space presents a much faster way of neural architectures search as well as the obtained calculation cells are able to be generalized for other tasks types. The following figure provides the NAS architecture used in ImageNet [7] dataset.

Architectures obtained by the NAS search space are in a feed-forward stocks. Each layer of the proposed neural network receives two inputs one direct input and a skip input from its previous layer. Layers provided are from two different categories: normal and reduction cells. The two types of cells are safe but the only difference between the two cells is that the reduction cell is followed by a stride of 2 in order to minimize the feature map dimensions while the normal cell keep the same input size.

Once the neural network architecture is specified, we have to specify two free parameters that are used during training N, and F while N is the number of normal cells and F is the number of convolution layers and output filters. In the proposed work, we set N=18 and F=448.

In order to develop our USCT image classification system, we used the aging evolutionary algorithm AmoebaNet-A. Models population will be initialized randomly by architectures from the NAS search space. The model that present the best calculation cells will be taken as a parent. A new architecture will be obtained using the mutation operation to obtain the child architecture. The child architecture is kept to be used to train and test the classification system. Child architectures are obtained by using three types of mutations:

- Hidden state mutation
- Identity mutation
- Op mutation

We note that when training the neural network architecture, one of these three mutations is applied randomly.

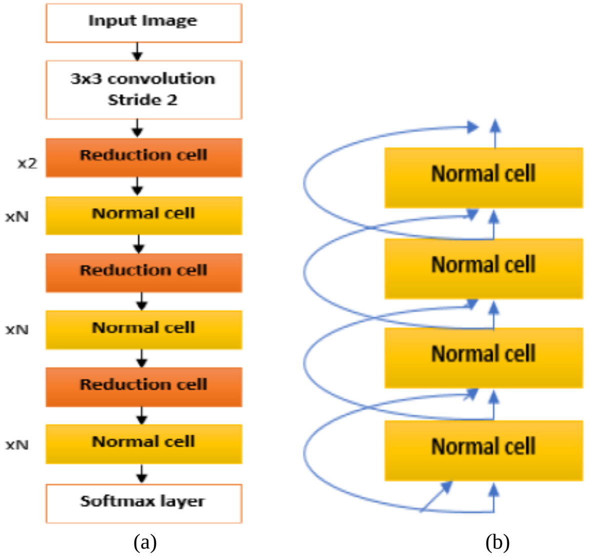


Fig 12. (a) NAS Search Space, (b) Detailed Architecture with Skip Input.

IV. Results

A. Train-test Accuracy Results

Due to transfer deep learning on our dataset, we have achieved 100% of accuracy in train process and 96%in test process classification accuracy as shown in Table III. More we are going deeper more the classification test accuracy is improved. We have achieved the top accuracy USCT image classification with Amoebanet Model by 96% of test-accuracy where Inception-V3 comes with the second top accuracy with 91.7%.

B. Histogram of Classification Results

Regarding obtained results as illustrated in the Figure.13 by the histogram, AmeobaNet with 32 layers has the top test accuracy against results accuracy given by Inception-V3 with its 42 layers, MobileNet with its 28 layers and NasNet with Deep learning Models. We conclude that when we are going deeper, more results accuracy is enhanced.

C. Time Process Results

Our framework is based on the python language of the Keras package and on an Nvidia Titan GPU using a Linux operating system. Graphics cards (GPU) are characterized by the large number of cores that the processors allow as well as very large memory integrated with these processors. They are very useful for several computer tasks, precisely for software implementations like deep learning algorithms.

TABLE III. Transfer Deep Learning Accuracy Results

Algorithm	Image Resolution	Train Accuracy	Test Accuracy
Mobile Net	256*256	100%	87.5%
NasNet	256*256	100%	75%
Ameobanet	256*256	100%	95.8%
SOM [4]	256*256	100%	94%
Inception-V3	256*256	100%	91.7%

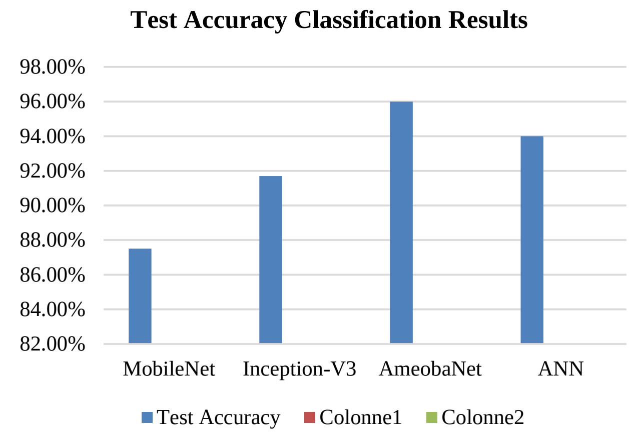


Fig 13. Histogram Results of Test-Accuracy USCT Images Classification.

Due to our implementation of our proposed neural networks models on GPU as shown by Annexes 1, 2, 3 and 4 in Annexes section , we have gained time process speed with 47.8 mn to each model trained on our USCT images against 149 mn with CPU models implementations as detailed by the following table.

TABLE IV. TIME-PROCESS RESULTS ON GPU AND CPU

Algorithm	CPU-Time- Process	GPU-Time-Process
MobileNet	146.9 mn	48mn
Inception-V3	145.52mn	47.6mn
AmeobaNet	148mn	49mn

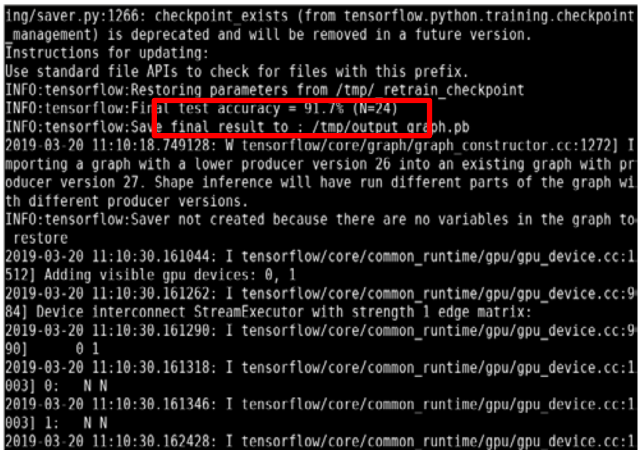
V. DISCUSSIONS

A comparative study was done with previously works, given results were shown in the Table. 5 below. AmeobaNet as an evolutionary deep neural network has achieved the best accuracy in our work and outperforms recent works [18, 19] with 22% of accuracy. Indeed, it is explained by the mind of the evolutionary algorithms which aims to search for the best architecture that can be used for a desired task. The architecture search is done in the network architecture search space (NAS). However, The Inception-V3 comes by a promising results accuracy with a value of 91.7% surpassing the state of the art with a test accuracy value equal to 11%. The inception v3 module computes multiple different transformations over the same input map in parallel, connecting the results into a single output. For each layer, it does a 5x5 convolution, 3x3 convolution, and max pooling, each carries different information, which of course is computationally costly. Therefore the authors of Inception decided to overcome this problem by introducing the dimension reductions. What is meant by dimension reduction is by using 1x1 convolution before going to the bottlenecks 3x3 and 5x5 convolutions. Therefore it has the compressed version of the spatial information. We have outperforming previously works. In Fact, we overcome [27] with inception –v3 with 13% and with SOM by 14% .With AmeobaNet we have overcome the major related works with neural networks classification models against Alex Net, NasNet, Inception V3 and MobileNet models accuracy classification [4, 5, 7, 24, 27].

TABLE V. COMPARATIVE STUDY WITH THE STATE OF THE ART

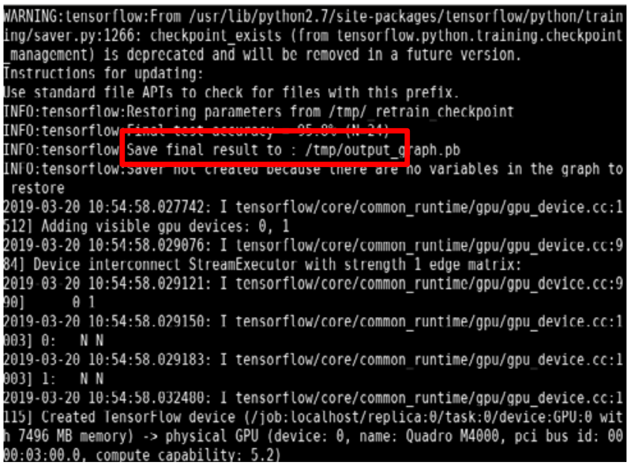
Algorithm	Our Work	Related Work
Mobil Net V2	87.5%	89.9% [24]
NasNet	75%%	80.8%
Ameobanet	**96%**	74,5 [21,22]
SOM [14]	94%	79% [3]
Inception-V3	91.7%	78.8% [27, 4, 5]

VI. ANNEXES



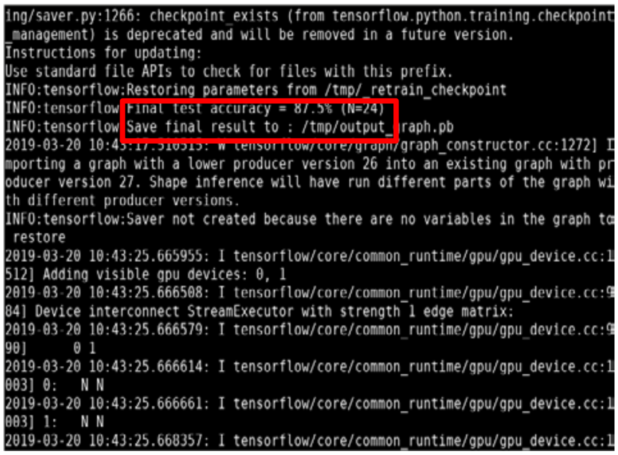
```
ing/saver.py:1266: checkpoint_exists (from tensorflow.python.training.checkpoint
_management) is deprecated and will be removed in a future version.
Instructions for updating:
Use standard file APIs to check for files with this prefix.
INFO:tensorflow:Restoring parameters from /tmp/_retrain_checkpoint
INFO:tensorflow:Final test accuracy = 91.7% (N=24)
INFO:tensorflow:Save final result to : /tmp/output_graph.pb
2019-03-20 11:10:18.749128: W tensorflow/core/graph/graph_constructor.cc:1272] I
mporting a graph with a lower producer version 26 into an existing graph with pr
oducer version 27. Shape inference will have run different parts of the graph wi
th different producer versions.
INFO:tensorflow:Saver not created because there are no variables in the graph to
 restore
2019-03-20 11:10:30.161044: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1
512] Adding visible gpu devices: 0, 1
2019-03-20 11:10:30.161262: I tensorflow/core/common_runtime/gpu/gpu_device.cc:9
84] Device interconnect StreamExecutor with strength 1 edge matrix:
2019-03-20 11:10:30.161290: I tensorflow/core/common_runtime/gpu/gpu_device.cc:9
90]      0 1
2019-03-20 11:10:30.161318: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1
003] 0:   N N
2019-03-20 11:10:30.161346: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1
003] 1:   N N
2019-03-20 11:10:30.162428: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1
```

Annex1. Inception V3 Transfer Deep Learning



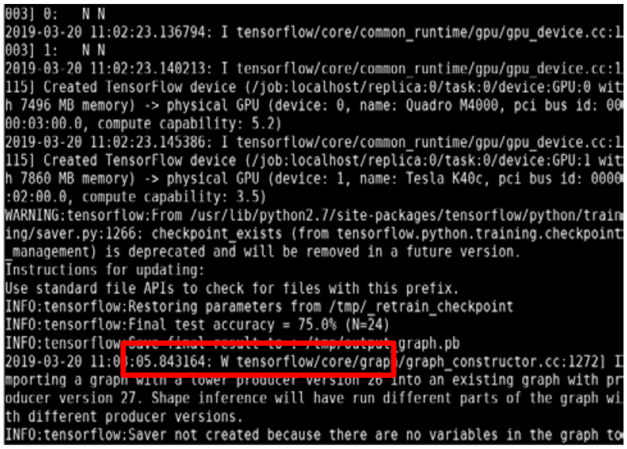
```
WARNING:tensorflow:From /usr/lib/python2.7/site-packages/tensorflow/python/train
ing/saver.py:1266: checkpoint_exists (from tensorflow.python.training.checkpoint
_management) is deprecated and will be removed in a future version.
Instructions for updating:
Use standard file APIs to check for files with this prefix.
INFO:tensorflow:Restoring parameters from /tmp/_retrain_checkpoint
INFO:tensorflow:Final test accuracy = 95.0% (N=24)
INFO:tensorflow:Save final result to : /tmp/output_graph.pb
INFO:tensorflow:Saver not created because there are no variables in the graph to
 restore
2019-03-20 10:54:58.027742: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1
512] Adding visible gpu devices: 0, 1
2019-03-20 10:54:58.029076: I tensorflow/core/common_runtime/gpu/gpu_device.cc:9
84] Device interconnect StreamExecutor with strength 1 edge matrix:
2019-03-20 10:54:58.029121: I tensorflow/core/common_runtime/gpu/gpu_device.cc:9
90]      0 1
2019-03-20 10:54:58.029150: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1
003] 0:   N N
2019-03-20 10:54:58.029183: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1
003] 1:   N N
2019-03-20 10:54:58.032480: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1
115] Created TensorFlow device (/job:localhost/replica:0/task:0/device:GPU:0 wit
h 7496 MB memory) -> physical GPU (device: 0, name: Quadro M4000, pci bus id: 00
00:03:00.0, compute capability: 5.2)
```

Annex2. AmeobaNet Transfer Learning



```
ing/saver.py:1266: checkpoint_exists (from tensorflow.python.training.checkpoint
_management) is deprecated and will be removed in a future version.
Instructions for updating:
Use standard file APIs to check for files with this prefix.
INFO:tensorflow:Restoring parameters from /tmp/_retrain_checkpoint
INFO:tensorflow:Final test accuracy = 87.5% (N=24)
INFO:tensorflow:Save final result to : /tmp/output_graph.pb
2019-03-20 10:43:17.510515: W tensorflow/core/graph/graph_constructor.cc:1272] I
mporting a graph with a lower producer version 26 into an existing graph with pr
oducer version 27. Shape inference will have run different parts of the graph wi
th different producer versions.
INFO:tensorflow:Saver not created because there are no variables in the graph to
 restore
2019-03-20 10:43:25.665955: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1
512] Adding visible gpu devices: 0, 1
2019-03-20 10:43:25.666508: I tensorflow/core/common_runtime/gpu/gpu_device.cc:9
84] Device interconnect StreamExecutor with strength 1 edge matrix:
2019-03-20 10:43:25.666579: I tensorflow/core/common_runtime/gpu/gpu_device.cc:9
90]      0 1
2019-03-20 10:43:25.666614: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1
003] 0:   N N
2019-03-20 10:43:25.666661: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1
003] 1:   N N
2019-03-20 10:43:25.668357: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1
```

Annex3. Mobile Net Transfer Learning



```
003] 0:   N N
2019-03-20 11:02:23.136794: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1
003] 1:   N N
2019-03-20 11:02:23.140213: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1
115] Created TensorFlow device (/job:localhost/replica:0/task:0/device:GPU:0 wit
h 7496 MB memory) -> physical GPU (device: 0, name: Quadro M4000, pci bus id: 00
00:03:00.0, compute capability: 5.2)
2019-03-20 11:02:23.145386: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1
115] Created TensorFlow device (/job:localhost/replica:0/task:0/device:GPU:1 wit
h 7860 MB memory) -> physical GPU (device: 1, name: Tesla K40c, pci bus id: 0000
:02:00.0, compute capability: 3.5)
WARNING:tensorflow:From /usr/lib/python2.7/site-packages/tensorflow/python/train
ing/saver.py:1266: checkpoint_exists (from tensorflow.python.training.checkpoint
_management) is deprecated and will be removed in a future version.
Instructions for updating:
Use standard file APIs to check for files with this prefix.
INFO:tensorflow:Restoring parameters from /tmp/_retrain_checkpoint
INFO:tensorflow:Final test accuracy = 75.0% (N=24)
INFO:tensorflow:Save final result to : /tmp/output_graph.pb
2019-03-20 11:03:05.843164: W tensorflow/core/graph/graph_constructor.cc:1272] I
mporting a graph with a lower producer version 26 into an existing graph with pr
oducer version 27. Shape inference will have run different parts of the graph wi
th different producer versions.
INFO:tensorflow:Saver not created because there are no variables in the graph to
```

Annex4. NasNet Transfer Learning Results

VII. Conclusion and Future Work

In this paper, a transfer Deep Learning method has been performed for USCT image classifications into three classes such as healthy, fractured and osteoporosis USCT images. Using inception-V3, MobileNet, and AmeobaNet, we have overcome previously works and has achieved excellent results in the area of medical image classification with a top value of 96% of test classification accuracy for a short time process. For future work, we have to think about automatic USCT image segmentation to detect different structure automatically using a variable model structure of neural network in first step [28]. Then using CNN models. Furthermore, a hardware implementation of our code will be done on Pynq-FPGA. Thus we will get a medical real-time application.

References

[1] Shi, Z., He, L., Suzuki, K., Nakamura, T., & Itoh, H. (2009). Survey on neural networks used for medical image processing. International journal of computational science, 3(1), 86.

[2] Review: NAS Net — Neural Architecture Search Network (Image Classification).

[3] Altaf, F., Islam, S., Akhtar, N., & Janjua, N. K. (2019). Going Deep in Medical Image Analysis: Concepts, Methods, Challenges and Future Directions. arXiv preprint arXiv:1902.05655.

[4] Y. Zhang, S. Chen, S. Wang, J. Yang, and P. Phillips, "Magnetic resonance brain image classification based on weighted-type fractional fourier transform and nonparallel support vector machine," International Journal of Imaging Systems and Technology, vol. 25, no. 4, pp. 317–327, 2015.

[5] ZhiFei Lai, Hui Fang Deng "Medical Image Classification Based on Deep Features Extracted by Deep Model and Statistic Feature Fusion with Multilayer. In: Perceptron"Compt Intell Neuroscience Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA) (pp. 19-23). IEEE.

[6] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, Gradient-based learning applied to document recognition, in Proceedings of the IEEE, vol. 86, no.11, pp. 2278-2324, Nov 1998

[7] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[8] Zheng, Y., Wang, H., & Hao, Y. (2020, April). CNN study of convolutional neural networks in classification and feature extraction applications. In Big Data II: Learning, Analytics, and Applications (Vol. 11395, p. 113950K). International Society for Optics and Photonics.

[9] Sae-Lim, W., Wettayaprasit, W., & Aiyarak, P. (2019, July). Convolutional Neural Networks Using MobileNet for Skin Lesion Classification. In 2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 242-247). IEEE.

[10] LIU, Yannan, WEI, Lingxiao, LUO, Bo, et al. Fault injection attack on deep neural network. In : 2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD). IEEE, 2017. p. 131-138.

[11] Sae-Lim, W., Wettayaprasit, W., & Aiyarak, P. (2019, July). Convolutional Neural Networks Using MobileNet for Skin Lesion Classification. In 2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 242-247). IEEE.

[12] Weng, Y., Zhou, T., Li, Y., & Qiu, X. (2019). Nas-unet: Neural architecture search for medical image segmentation. IEEE Access, 7, 44247-44257.

[13] Lai, Z., & Deng, H. (2018). Medical Image Classification Based on Deep Features Extracted by Deep Model and Statistic Feature Fusion with Multilayer Perceptron. Computational intelligence and neuroscience, 2018.

[14] Fradi, M., Lasaygues, P., & Machhout, M. (2019, March). Auto-Organiser Neural Network Application for Ultrasound Computed Tomographic Image Classification. In 2019 19th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA) (pp. 19-23). IEEE.

[15] Mehdy, M. M., Ng, P. Y., Shair, E. F., Saleh, N. I., & Gomes, C. (2017). Artificial neural networks in image processing for early detection of breast cancer. Computational and mathematical methods in medicine, 2017.

[16] Ruiter, N. V., Zapf, M., Hopp, T., Gemmeke, H., & van Dongen, K. W. (2017, March). USCT data challenge. In Medical Imaging 2017: Ultrasonic Imaging and Tomography (Vol. 10139, p. 101391N). International Society for Optics and Photonics.

[17] Kaur, P., Sharma, P., & Palmia, A. (2020). Fuzzy clustering-based image segmentation techniques used to segment magnetic resonance imaging/computed tomography scan brain tissues: Comparative analysis. International Journal of Imaging Systems and Technology.

[18] Kaur, P., & Chaira, T. (2020). A novel fuzzy approach for segmenting medical images. Soft Computing, 1-11.

[19] Marwa Fradi, Wajih Elhadj Youssef, Ghaith Bouallegue, Mohsen Machhout and Philippe Lasaygues "Automatic USCT Image Processing Segmentation for Osteoporosis Detection" Springer Nature Switzerland AG 2020M. S. Bouhlel and S. Rovetta (Eds.): SETIT 2018, SIST 146, pp. 372–381, 2020.https://doi.org/10.1007/978-3-030-21005-2_36

[20] Fradi, M., Youssef, W. E., Lasaygues, P., & Machhout, M. (2018). Improved USCT of paired bones using wavelet-based image processing. International Journal of Image, Graphics and Signal Processing, 10(9), 1

[21] Cai, H., Zhu, L., & Han, S. (2018). Proxylessnas: Direct neural architecture search on target task and hardware. arXiv preprint arXiv:1812.00332

[22] Real, E., Aggarwal, A., Huang, Y., and Le, Q. V. Regularized evolution for image classifier architecture search. AAAI, 2019

[23] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861

[24] SANDLER, Mark, HOWARD, Andrew, ZHU, Menglong, et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 4510-4520

[25] SLOSS, Andrew N. et GUSTAFSON, Steven. 2019 Evolutionary Algorithms Review. arXiv preprint arXiv:1906.08870, 2019.

[26] B.Zoph,V.Vasudevan,J.Shlens,andQ.V.Le. Learning transferable architectures for scalable image recognition. In CVPR, 2018.

[27] Nguyen, L. D., Lin, D., Lin, Z., & Cao, J. (2018). Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation. 2018 IEEE International Symposium

[28] Bouallegue, K. (2017). A new class of neural networks and its applications. Neurocomputing, 249, 28-47.

Ice Concentration Estimation Method with Satellite based Microwave Radiometer by Means of Inversion Theory

Kohei Arai
Faculty of Science and Engineering
Saga University, Saga City, Japan

***Abstract*—Ice concentration estimation method with satellite-based microwave radiometer by means of inversion theory is proposed. Through experiments, it is found that the proposed methods are superior to the existing methods, the NASA Team algorithm and the Comiso's Bootstrap algorithm with up to 45% of improvement on ice concentration estimation accuracy based on the simulation study. Also 1.5 to 2.1% of improvement was achieved for the proposed method compared to the NASA Team and Comiso's Bootstrap algorithms for the actual The Special Sensor Microwave Imager (SSM/I) data of Okhotsk using Japanese Earth Resources Satellite: JERS-1/Synthetic Aperture Radar: SAR data as a truth data for estimating ice concentration.**

Keywords—Ice concentration; Microwave radiometer; Inversion Theory; Comiso's Bootstrap algorithm; The Special Sensor Microwave Imager (SSM/I); Japanese Earth Resources Satellite: JERS-1; Synthetic Aperture Radar: SAR

I. INTRODUCTION

Observation of sea ice is important in considering the global environment. This is because the sea ice areas on both poles of the earth cover 10% of the sea surface and not only have a great influence on the heat balance of the earth, the movement of the ocean and the atmosphere, but are also the places where the effects of global warming phenomena are likely to appear [1]. In addition to scientific observation purposes, it also has the purpose of preventing marine accidents, such as ensuring navigation and sea work safety in polar regions and ice floes.

Global maps of sea ice concentration, age and surface temperature derived from NIMBUS-7 satellite onboard Special Sensor Microwave Radiometer: SSMR: A case study is conducted and well reported together with importance of ice concentration estimation for global warming [2]. A microwave technique for mapping thin ice is investigated [3].

Special Sensor Microwave Imager: SSM/I concentrations using the bootstrap algorithm as the NASA standard product is well documented [4]. Also, temperature corrected bootstrap algorithm is proposed [5].

Estimating occupancy of in-pixel covering class by solving inverse problem, is conducted and well reported [6]. Area ratio estimation (mixing ratio estimation) by pixel category decomposition is proposed and validated [7]. Category decomposition based on maximum likelihood estimation is proposed [8].

Advanced Microwave Scanning Radiometer: AMSR is well reported in terms of requirements and preliminary design study [9]. Method for proportion estimation of mixed pixels by means of inversion problem solving is proposed for mixing ratio estimation [10].

On the other hand, inversion techniques for proportion estimation of Mixed Pixel: Mixels in high spatial resolution of satellite image is proposed [11]. Inversion for emissivity-temperature separation with Advanced Spaceborne based Sensor for Thermal Emission and Radiation: ASTER data is proposed [12]. Meanwhile, method for ice concentration estimation with microwave scanning radiometer data by means of inversion is proposed [13].

Estimation accuracy of ice concentration is not good enough for the global warming problem-solving. Therefore, strong demands of improvement of ice concentration estimation accuracy are raised among the global change research community. In this paper, a microwave radiometer installed on an artificial satellite is used to observe sea ice, and the method of category decomposition is used based on the brightness temperature data of multiple frequencies observed by the microwave radiometer. The author proposes a method to estimate ice concentration.

In the following section, related research works are described. Then, the proposed method is described followed by experimental set-up together with experimental results. After that, concluding remarks and some discussions are described.

II. RELATED RESEARCH WORKS

A method for ice concentration estimation with microwave radiometer data by means of inversion techniques is proposed [14]. An inversion for emissivity-temperature separation with ASTER data is proposed [15].

Spatial resolution enhancement by means of inversion is proposed [16]. Inversion techniques for proportion estimation of Mixels in high spatial resolution of satellite image analysis is also proposed [17].

Ice concentration estimation based on local inversion is proposed [18]. Application of inversion theory for image analysis and classification is investigated [19]. Sea Surface Temperature: SST estimation method with linearized inversion of Radiative Transfer Equation: RTE code for Advanced Earth Observing Satellite: ADEOS / Ocean Color and Temperature Scanner: OCTS is proposed [20].

Estimation of SST, wind speed and water vapor with microwave radiometer data based on simulated annealing is proposed as one of the microwave radiometer data applications [21]. Nonlinear optimization-based SST estimation methods with remote sensing satellite-based Microwave Scanning Radiometer: MSR data is proposed. [22].

Simultaneous estimation of geophysical parameters with microwave radiometer data based on accelerated Simulated Annealing: SA is proposed [23]. Meanwhile, sensitivity analysis for water vapor profile estimation with infrared sounder data based on inversion is proposed [24].

Data fusion between microwave and thermal infrared radiometer data and its application to skin sea surface temperature, wind speed and salinity retrievals is proposed [25]. Comparative study of optimization methods for estimation of SST and ocean wind with microwave radiometer data is proposed [26].

III. Conventional Method

A. Traditional Method for Estimation of Sea Ice Concentration

The estimation of sea ice concentration using a microwave radiometer has been actively performed since 1972, when the satellite NIMBUS-5 equipped with the microwave radiometer Electrically Scanning Microwave Radiometer: ESMR was launched. Gloersen et al. Have proposed a method for estimating the sea ice concentration of Antarctic one-year ice by the following equation.

$$C=(Tb-135)/(\varepsilon Ts-135) \quad (1)$$

where, C is the sea ice concentration, Tb is the brightness temperature observed by the sensor, Ts is the physical temperature of the 1-year ice, ε is the emissivity of the ice, and is 0.92 in the case of the 1-year ice in the nadir. The constant 135 is the sum of the brightness temperature of the open water surface (120K) and the atmospheric radiation (15K).

U.S. Navy uses the same NIMBUS-5 / ESMR, and as shown in Fig. 1, the brightness temperature when sea ice is 100% is 240K, and the brightness temperature when seawater is 100% is 135K, and the observed brightness temperature during that period is 135K.

The sea ice concentration was expressed as a linear relationship with the brightness temperature, and the sea ice concentration map was created constantly. In Japan, MOS-1 / MSR is often used to estimate the sea ice concentration of one-year ice. For example, a formula using a band of 31 GHz has been proposed as follows. Here, is the output (digital number) D of the 31 GHz channel.

$$C = 4.17D-220.83 \quad (2)$$

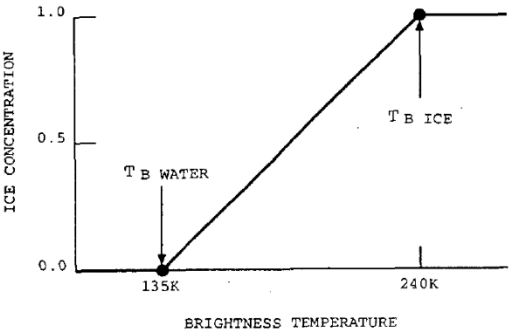


Fig. 1. A Relationship between Ice Concentration and Brightness Temperature.

B. Comiso's Bootstrap Algorithm

Furthermore, recently, NASA Team algorithm and Comiso's Bootstrap algorithm have been proposed to improve the estimation accuracy [3],[4]. In the NASA Team algorithm, a linear combination function of the sum / difference ratio (PR) of the vertical and horizontal polarization channels at 19 GHz and the sum / difference ratio (GR) of the vertically polarized waves at 37 GHz and 19 GHz is used by regression analysis beforehand. It is estimated using the obtained coefficients. At this time, it is considered that it is possible to estimate the composition ratio separately for one-year ice and perennial ice, and it is also called a weather filter to remove cloud and water vapor. It is also being done.

$$PR=(Tb(19V)-Tb(19H))/(Tb(19V)+Tb(19H)) \quad (3)$$

$$GR=(Tb(37V)-Tb(19V))/(Tb(37V)+Tb(19V)) \quad (4)$$

$$f=(C_1+C_2PR+C_3GR+C_4PR*GR)/D \quad (5)$$

$$C_m=(C_9+C_{10}PR+C_{11}GR+C_{12}PR*GR)/D \quad (6)$$

$$D=C_5+C_6PR+C_7GR+C_8PR*GR \quad (7)$$

where, the Weather Filter means that if the GR calculated from 19 and 37 GHz is 0.05 and the GR calculated from 19 and 22 GHz is 0.045 or more, the sea ice concentration is 0.

Here, C_f and C_m are the composition ratios of one-year and perennial ice. In Comiso's Bootstrap algorithm, training samples are pre-instructed by a human in advance, and they are referred to obtain the sea ice concentration by the following formula.

$$IC=\frac{\sqrt{(T_{b1}-T_{b1}^{W})^2+(T_{b2}-T_{b2}^{W})^2}}{\sqrt{(T_{b1}^{l}-T_{b1}^{W})^2+(T_{b2}^{l}-T_{b2}^{W})^2}} \quad (8)$$

where, *IC* is the sea ice density, T_{bx} is the brightness temperature of different frequency channels (19 and 37 GHz are often used), and $T_{bx}{}^{w}$ is the open surface and ice brightness temperature, respectively. Represents temperature. Furthermore, in the Comiso's Bootstrap algorithm, a sea surface mask algorithm is also proposed, and correction of sea ice temperature is also considered [5].

Thus, several formulas for estimating sea ice concentration have been proposed, but it is said that the antenna brightness temperature of a microwave radiometer is expressed as a linear combination by the mixing ratio of the brightness temperature of sea ice and sea water. It is based on assumptions. However, as can be seen from the two-dimensional scatter diagram of vertically polarized waves of 19 and 37 GHz of SSM / I in the Arctic Ocean region (image shown in Fig. 3) on January 1, 1989, as shown in Fig. 2.

The distributions of 1-year ice and perennial ice have a large overlap and show a large dispersion, and further, they also largely overlap with those of cloud or water vapor-rich regions. Therefore, it is difficult to improve the estimation accuracy of sea ice concentration.

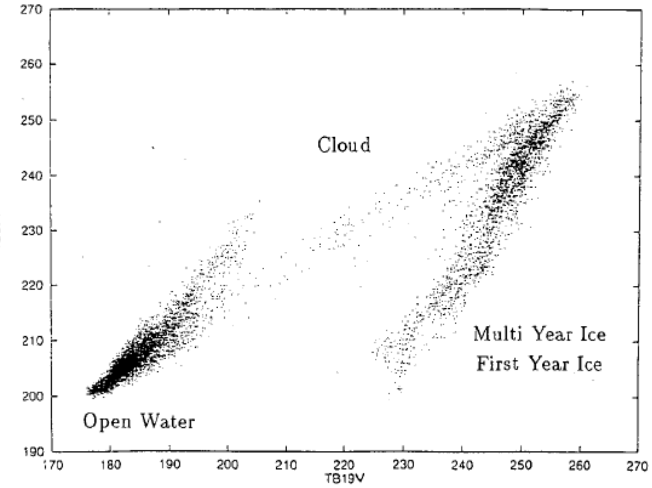


Fig. 2. Two-Dimensional Scatter Plot between 19 V and 37 V Channels of SSM/I Data of Arctic Ocean Area Acquired on Jan. 1 1989.

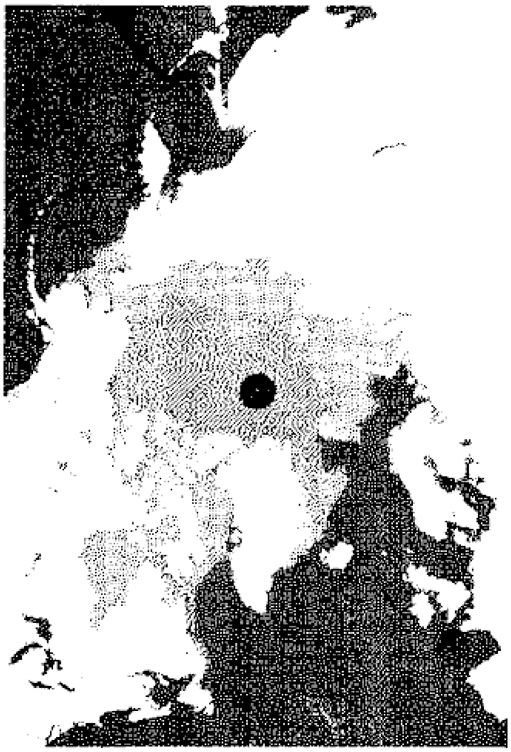


Fig. 3. A Portion of SSM/I 19 GHz H-Polarization of Image.

In addition, the antenna brightness temperature of the microwave radiometer changes depending on other factors such as water vapor and cloud water content. Therefore, an algorithm that considers those influences is necessary. In this research, we tried to use the method of category decomposition for sea ice concentration estimation. In doing so, the effects of cloud and water vapor are added by adding them to the category.

IV. Proposed Method

A. Linear Model of Pixel Spectral Vector

The author has already proposed a method of category decomposition based on the solution of the inverse problem6). According to this, when k types of categories are included in the pixels of the multiple spectral image, if the area ratio of each category is a_j, (j = 1, ..., k), the spectral vector P is It is represented by the following equation by a typical spectral vector M_j.

$$P = \sum_{j=1}^{k} a_j M_j \qquad (9)$$

$$\sum_{j=1}^{k} a_j = 1 \qquad (10)$$

$$a_j \geq 0 \qquad (11)$$

The area ratio can be estimated by the following equation, where n is the number of observation channels and k is the number of decomposed categories.

P=MA

$$P = [p_1, p_2, \dots, p_n]^t$$

$$M = [M_1, M_2, \dots, M_k]$$

$$A = [a_1, a_2, \dots, a_k]^t \qquad (12)$$

P is a vector that aligns the brightness temperatures observed at different frequencies, and M_j is a vector that consists of representative brightness temperatures of categories such as sea ice and open water surface in each observation channel. A is a vector showing the area ratio of the category. Hereinafter, P, A, M_j, and M will be referred to as an observation vector, an area ratio vector, an average vector, and an average matrix, respectively.

It is already known that M can be estimated by extracting training samples from the image, and P is an observed value. Based on this basic model, the area ratio of each category (one of them is sea ice concentration) is obtained by using the inverse problem-solving method. This method includes the Moore-Penrose type generalized inverse matrix, the observed least squares method, the area ratio least squares method, the maximum likelihood grid search method, etc., as shown in [6].

B. Moore-Penrose Type Generalized Inverse Matrix

In the above linear model, if the mean matrix M is regular, there is an inverse matrix:

$$A = M^{-1}P \qquad (13)$$

Therefore, the area ratio can be easily calculated. However, since M is rarely regular, the Moore-Penrose generalized inverse matrix shown below is needed.

$$A = (M^tM)^{-1}M^tP = M^+P \tag{14}$$

However, since this method does not consider the measurement error of the data, the estimation error may be large, or the total of the mixing ratios may not be 1. Therefore, the method using the least squares method with the following constraint conditions is used.

C. Estimation using the Least Squares Method based on Observed Value

If the observed value contains an error, the accuracy is improved by minimizing the error between the observed vector P and the estimated value $P' = MA$ of the true value. The restraint conditions at that time are as follows:

$$|P\text{-}MA| \rightarrow \text{min}$$

$$u^tA = 1\ (u = [1,1,\dots,1]^t)$$

$$a_j \geq 0 (j = 1,2,\dots,k) \tag{15}$$

If the non-negative condition a_j=0 is removed, it can be analytically solved by using Lagrange's undetermined multiplier method as the constrained least squares method. Therefore, let λ be an undetermined multiplier and consider the following function F.

$$F(A,\lambda) = \frac{1}{2}|P - MA|^2 - \lambda(u^tA - 1) \tag{16}$$

A can be obtained by solving the following simultaneous equations. Eq. (17) can be rewritten in vector form and M^t $(P - MA) - \lambda u = 0$.

$$\frac{\partial F}{\partial a_j} = -m_j^t(P - MA) - \lambda = 0(j = 1,2,3) \tag{17}$$

$$\frac{\partial F}{\partial a_j} = -(u^tA - 1) = 0 \tag{18}$$

This equation can be solved for A if m_j (1,2,3) is first-order independent, and becomes.

$$A = M^+P + \lambda (M^tM)^{-1}u \tag{19}$$

Substituting this into Eq. (18) gives and λ is determined.

$$\lambda = \frac{1-u^tM^+P}{u^t(M^tM)^{-1}u}(M^tM)^{-1}u \tag{20}$$

Therefore, from Eq. (19), the estimated value of the area ratio vector A is obtained as.

$$A = M^+P + \frac{1-u^tM^+P}{u^t(M^tM)^{-1}u}(M^tM)^{-1}u \tag{21}$$

By the way, in the actual measurement data, the matrix M is not always equal to the true representative value vector M_o. Let this difference be E. If the true area ratio vector A_0 and the observation vector does not include an error, the following equation holds.

$$M=M_0+E$$

$$P=M_0A_0 \tag{22}$$

Therefore, when M and P are given by Eq. (22), the area ratio vector A_1 obtained by Eq. (21) is

$$A_1 = A_0 - M^+EA_0 + \frac{u^tM^+EA_0}{u^t(M^tM)^{-1}u}(M^tM)^{-1}u \tag{23}$$

which is an estimated error of the area ratio vector can be estimated by the following Eq. (24).

$$A_1 - A_0 = - M^+EA_0 + \frac{u^tM^+EA_0}{u^t(M^tM)^{-1}u}(M^tM)^{-1}u$$

$$\geq \left[\left\{\frac{u^tM^+EA_0}{u^t(M^tM)^{-1}u}(M^tM)^{-1}u\right\} - \{M^+EA_0\}\right]$$

$$= \left[\frac{cos(\beta)}{cos(\alpha)} - 1\right]|M^+EA_0| \tag{24}$$

where α is the angle between u and $(M^tM)^{-1}u$, and β is the angle between u and M^+EA_0. On the other hand, the area ratio vector A' calculated algebraically for the same data and its estimation error are as follows:

$$A' = A_0 - EA_0$$

$$|A' - A_0| = |M^+EA_0| \tag{25}$$

and from Eq. (24), in some cases a much larger error can occur than in the algebraic solution.

That is, the estimation accuracy depends on the accuracy of the matrix M composed of the representative values of each category.

D. Estimation by Area Ratio Least Squares Method [7]

In the least-squares method of observed values, the area ratio was obtained because the problem of the representativeness of M can be expressed as the error of the observation vector. However, when using the actual measurement data, the representativeness of M is more important than P, and it has a great influence on the estimation accuracy of the area ratio. Therefore, the estimated value N of the generalized inverse matrix M_0^+ of the true representative value matrix and the area ratio vector A are set as unvalued, and A that minimizes the residual difference between M^+ and N is obtained from the following formula.

$$|M^+\text{-}N| \rightarrow \text{min}$$

$$A=NP$$

$$u^tA=1 \tag{26}$$

The area ratio vector obtained by this method agrees with the one obtained by the following least squares method with constraints.

$$|A\text{-}M^+P| \rightarrow \text{min}$$

$$u^tA=1 \tag{27}$$

bNow, again, Lagrange's undetermined multiplier method is used to solve equation (27). Eliminating A from Eq. (27) from the residual $V = M^+\text{-}N$, the following equation is obtained.

$$|V| \rightarrow \text{min}$$

$$U^t(M^+\text{-}V)P=1 \tag{28}$$

Let λ be an undetermined multiplier and consider the following function.

$$F(V,\lambda) = \frac{1}{2}|V|^2 - \lambda(u^t(M^+ - V)P - 1) \qquad (29)$$

V is obtained as the solution of the simultaneous cubic equation (Eq. (30), (31).

$$\frac{\partial F}{\partial V_{ji}} = V_{ji} + \lambda p_i = 0\ (j = 1,2,..,n),(i = 1,2,...,n) \qquad (30)$$

$$\frac{\partial F}{\partial \lambda} = -\{u^t(M^+ - V)P - 1\} \quad = 0 \qquad (31)$$

Also, Eq. (30) can be rewritten in the form of a matrix and becomes m.

$$V=- \lambda u P^+ \qquad (32)$$

By substituting this into Eq. (31) and determining λ,

$$\lambda = \frac{1-u^tM^+P}{|u|^2|P|^2} \qquad (33)$$

Therefore, *V*, *N*, and *A* are obtained as follows.

$$V = -\frac{1-u^tM^+P}{|u|^2|P|^2}uP^2 \qquad (34)$$

$$N = M^+ + \frac{1-u^tM^+P}{|u|^2|P|^2}uP^t \qquad (35)$$

$$A = M^+P + \frac{1-u^tM^+P}{|u|^2}u \qquad (36)$$

where, if *M* and *P* are given in the same way as in the case of the observed least squares method, the estimated area ratio vector A_2 and the estimation error | A_2-A_0 | are given by (37) and (38).

$$A_2 = A_0 - M^+EA_0 + \frac{1-u^tM^+P}{|u|^2}u \qquad (37)$$

$$|A_2 - A_0|^2 = |M^+EA_0|^2 - \frac{1-u^tM^+EA_0}{|u|^2} \qquad (38)$$

The estimation accuracy of the area ratio least squares method is as good as the second term on the right side of (38), and the problem that the estimation accuracy is extremely poor depending on the property of the matrix *M* as in the conventional method is solved. It seems to be that. $u^tA = 1$ is an equation of a plane in k-dimensional vector space, where *A* is a variable, and its normal vector is *u*. M^+P is also a point in the k-dimensional vector space.

This is to find the point on the plane u^tA_1 with the minimum distance from the point M^+ *P*. In other words, it is enough to find the foot of the perpendicular line from the point M^+ *P* to the plane $u^tA = 1$. Since the perpendicular is parallel to the vector u, the equation of the perpendicular can be written as *v* with λ as a parameter.

$$A=M^+P+ \lambda u \qquad (39)$$

Eq. (36) is obtained by finding the intersection of this and the plane.

$$u^tA = 1 \qquad (40)$$

then eliminating *A* and finding λ.

E. Maximum Likelihood Grid Search Method [8]

As in the case of the least square method, the observation value vector is *P*, the area ratio vector is *A*, and the representative value matrix is *M*. At that time, it is assumed that the observation value is given in a form in which the observation error ε is added to the linear combination of *M* and *A*.

$$P=MA+ \varepsilon$$

$$P = [p_1, p_2, \dots, p_n]^t$$

$$M = \begin{pmatrix} m_{11} & \dots & m_{1k} \\ \vdots & \vdots & \vdots \\ m_{n1} & \dots & m_{nk} \end{pmatrix}$$

$$A = [a_1, a_2, \dots, a_k]^t$$

$$\varepsilon = [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_k]^t \qquad (41)$$

Then, assume that the element m_{ij} protection of *M* follows the normal distribution: $N(m_{ij}^*, \sigma_{ij}^2)$ with mean m_{ij} and variance $\sigma_{ij}{}^2$, and ε_i follows $N(o, \sigma_{ei}^2)$, and consider the observation value vector *P* as a random variable. At that time, the observed value of the i-th band pi follows the normal distribution $N(m_i^*, \sigma_i^2)$ of mean mi and variance $\sigma_{ij}{}^2$ expressed by the following equations (42) and (43). However, the representative values of each category are assumed to be independent of each other.

$$m_i = m_i^* \cdot A, m_i^* = [m_{i1}^*, m_{i2}^*, \dots, m_{ik}^*] \qquad (42)$$

$$\sigma_i^2 = A^t \cdot S_i + \sigma_{ei}^2, S_i = diag(\sigma_{i1}^2, \sigma_{i2}^2, \dots, \sigma_{ik}^2) \qquad (43)$$

where diag represents a diagonal matrix. The probability *Q* (p_i) that the observed value p_i of the i-th band is observed is expressed by the following equation.

$$Q(p_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} exp\left\{-\frac{(p_i-m_i)^2}{2\sigma_i^2}\right\} \qquad (44)$$

where the probability that the observed value vector *P* is observed is *Q* (*P*), it is expressed by the following equation.

$$Q(P) = \Pi_{i=1}^n Q(p_i) \qquad (45)$$

The calculated area ratio *A* is the value when *Q* (*P*) is at its maximum. Therefore, the area ratio *A* that minimizes the following equation *R* (*P*) must be obtained.

$$R(P)=-ln\{Q(P)\} \qquad (46)$$

$$\sum_{j=1}^k A_j = 1, A_j \geq 0, (j = 1,2,\dots,k) \qquad (47)$$

In practice, the ratio of each category is changed every 1%, *R* (*P*) is calculated for all the combinations, and *A* is determined. Therefore, it is called the maximum likelihood grid search method.

V. Experiments

The most important thing in valid intercomparing between methods is correct data. Here, we tried the evaluation by the simulation data that can give the correct answer and the evaluation by the correct answer data created from the synthetic aperture radar image.

A. Simulation Data Used

The SSM / 1 data of 19GHz vertical, horizontal polarization, 22GHz vertical polarization, 37GHz vertical, horizontal polarization of the above-mentioned January 1, 1989 Arctic were used. Simulation data is created based on the training sample data extracted from SSM / I data, and the proposed method is applied to compare the accuracy. The procedure for creating the simulation data was as follows.

1) Create mixture ratio data for multiple categories.

2) Extract training samples for the category to be classified from the SSM / I image.

3) Considering the variance of the training sample in (2), output the data given the error according to the normal distribution according to the mixing ratio in (1) for the number of bands.

4) Obtain a new training sample from the data obtained in (3).

In this experiment, there are four categories: annual ice, perennial ice, open water surface, and clouds, and sea ice concentration is calculated as the sum of the mixing ratios of annual ice and perennial ice. 1-year ice is from the mouth of the Amur River to the open ocean, perennial ice is from the Arctic, open water is from the Scandinavian Peninsula and the North Atlantic Ocean in Greenland, and clouds are in the Newfoundland and central Portugal. From the located Atlantic Ocean, respectively, they were extracted as training samples. The bands used are 5 bands of 19H, 19V, 22V, 37H, 37V. Table I is a training sample of each category extracted from the actual SSM / I image.

B. Simulation Result

Table II shows the Root Mean Square: RMS error and CPU Time (Elapsed time) from the correct data.

In this table, LSQ is the least squares and MLH is the maximum likelihood grid search method, respectively. As for the inverse problem-solving method, the method with the constraint condition is more accurate. It is also slightly better than the NASA Team and Comiso's Bootstrap algorithms. Furthermore, the maximum likelihood grid search method was confirmed to improve the accuracy by 45% with respect to the Comiso's Bootstrap algorithm. However, as the number of categories increases, the amount of calculation increases exponentially, so it takes considerably longer than the linear solution method.

C. Experimental Method with Real SAR Data

SAR has a much higher spatial resolution than SSM / I. In addition, in the case of L-band SAR, sea ice generally has a higher backscattering cross section than the sea discharge surface, so it is relatively easy to identify it if effects such as sea surface wind speed are taken into consideration. Furthermore, because of its long wavelength, it is hardly affected by clouds. Utilizing this, sea ice concentration data corresponding to SSM / I images was created from SAR images, and the accuracy of the inverse problem-solving method was evaluated using the data as correct answer data.

First, for each pixel of the SAR image, the categories of sea ice and open water surface are classified using the difference in backscattering cross section. Due to the difference in resolution, there are about 7,000 SAR pixels corresponding to SSM / I pixels. Since the latitude and longitude are known in both images, the sea ice concentration was estimated from the SAR image by matching both images based on them.

JERS-1 / SAR images of the Sea of Okhotsk facing the coastline of northern Hokkaido were used. The date and time are from February 4 to 9, 1994, and the exact location is 143~146 degrees east longitude and 43.5~46 degrees north latitude. One pixel of JERS-1 is 12.5mx 12.5m, but this time we used the data processed to 300mx 300m. Each pixel is data corresponding to −20 dB to 5.5 dB. Here, the important point is the threshold that distinguishes sea ice from open water. Fig. 4 is a histogram of pixel values of the SAR image used this time.

The higher backscattering cross section is sea ice, and the lower backscatter cross section is the open water surface distribution. From this figure, it is appropriate to set it to 60 to 70 (actual value is -14dB to -13dB). In the experiment, -13.5 dB was used as the threshold.

Fig. 5 shows the image of JERS-1/SAR data of Okhotsk acquired on February 4 to 9 1994.

TABLE I. MEAN AND VARIANCE OF THE TRAINING SAMPLES

	Mean of Brightness Temperature (K)				
	19H	19V	22V	37H	37V
Multi Year Ice	209.8	228.2	224.4	193.8	206.8
First Year Ice	235.1	246.4	244.3	229.4	236.7
Open Water	105.1	179.4	187.8	138.1	203.6
Could	163	211.4	235.3	208.6	237.7
	Standard Deviation of Brightness Temp. (K)				
Multi Year Ice	24.82	23.19	23.9	35.57	40.86
First Year Ice	25.3	28.24	28.01	34.3	28.45
Open Water	24.26	12.31	15.2	33.58	12.72
Could	120.9	69.22	76.75	188.37	99.68

TABLE II. A COMPARISON OF RMS ERROR AND CPU TIME

Method	RMSE	CPU Time(sec)
Moore-Penrose	7.83	4.32
LSQ Minimizing Observation Vector	7.66	4.43
LSQ Minimizing Mixing Vector	7.53	4.17
MLH Search All Possible Solution	4.2	9.88
NASA Team Algorithm	7.71	4.26
COMISO's Bootstrap Algorithm	7.67	4.75

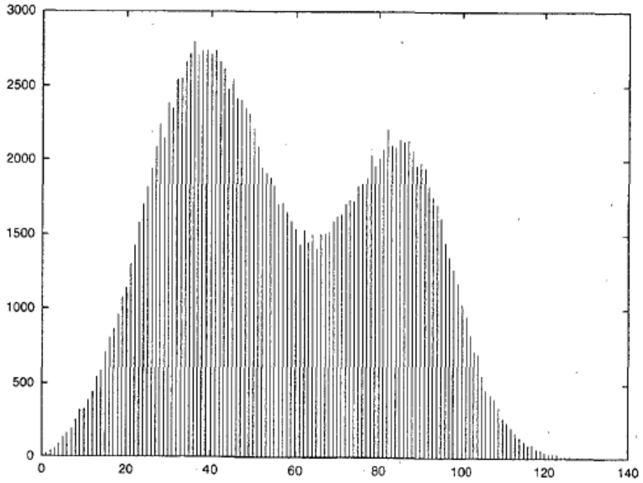


Fig. 4. Histogram of the SAR Image (Digital Number VS Frequency).

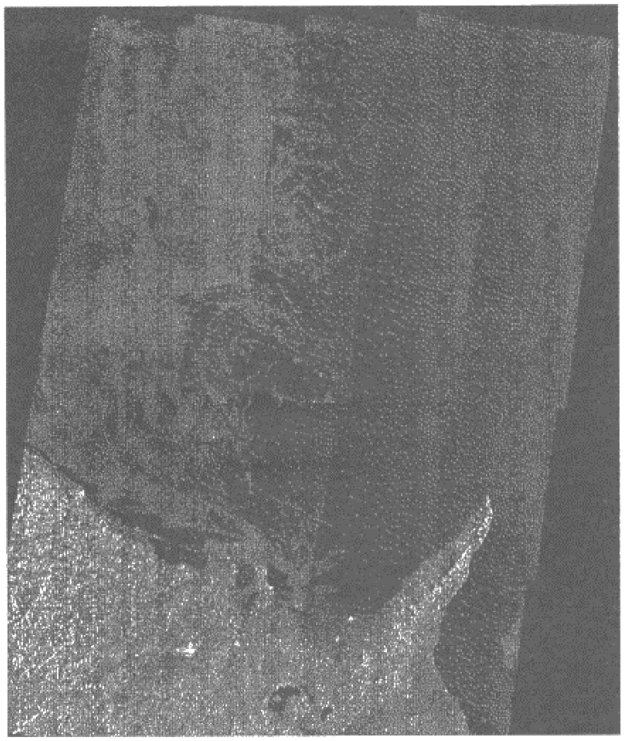


Fig. 5. JERS-1/SAR Image of Okhotsk acquired on February 4 to 9, 1994.

D. Experimental Result with Real SAR Data

Looking at the RMS error of each method, Moore-Penrose generalized inverse matrix (13.63%), observed least squares (12.96%), area ratio least squares (12.80%), NASA algorithm (12.66%), Comiso's Bootstrap algorithm (12.58%) and the maximum likelihood grid search method (12.40%) were confirmed, and it was confirmed that the accuracy was high in almost the same order as the comparative evaluation results by simulation. However, a low estimation accuracy was obtained overall, which is considered to be due to the following reasons.

1) Error in alignment
2) Error of estimated data from SAR image
3) Moving sea ice

Due to the large difference in the resolution of the images used this time, 25 km square and 0.3 km square, and the physical characteristics of sea ice concentration, it is possible that a slight error in the registration caused a large error in the estimation results.

The author artificially generated images that were shifted by 0.5 pixels in SMM / I and in the north, south, east, and west. From the result, there is a possibility that the maximum is shifted by 0.5 pixels. If SSM / I pixel shifts by 0.5 pixel, it is 12.5km, which is quite large. In addition, since the sea ice present at this site is drift ice, it is considered that the difference between the observation times of both data is also relevant. Furthermore, as can be seen from the histogram, the distribution of backscatter values of sea ice and the distribution of objects on the sea surface largely overlap. An estimation error occurs from here. The threshold that minimizes the estimation error can be set but cannot be 0. The reason for this distribution is that the surface condition of the drift ice is complicated and has various forms.

VI. Conclusion

From the result of the simulation experiment and the experiment using the sea ice concentration estimated from the synthetic aperture radar, it was confirmed that the inverse problem solution is effective for the sea ice concentration. In particular, the least-squares method with the constraint of the square of the estimation error of the observed value and the estimation error of the mixture ratio was found to be effective. However, the accuracy of estimation is slightly inferior to that of the maximum likelihood grid search method because the calculated area ratio does not include a non-negative condition and the variance of the brightness temperature of the training samples in the category is not taken into consideration.

Compared to these, the NASA Team and Comiso's Bootstrap algorithms have almost the same estimation accuracy as the least-squares method that constrains the squares of the estimation error of the observed value and the estimation error of the mixture ratio, and the maximum likelihood grid search method. It was also confirmed that the estimation accuracy was slightly inferior.

The maximum likelihood grid search method was 2.1% and 1.5% higher than that of NASA Team and Comiso's Bootstrap algorithm, respectively, based on the results of experiments using the correct solution of sea ice concentration estimated from the synthetic aperture radar in the Okhotsk sea area. It was confirmed that the estimation accuracy of was improved. The reason is that only the maximum likelihood grid search method considers not only the average value of the brightness temperature of each category but also the variance.

VII. Future Research Works

The proposed method must be tested with the other microwave radiometer data as well as synthetic aperture data. Also, influences of sea ice concentration on global warming based on the estimated concentrations.

ACKNOWLEDGMENT

Lastly, the author would like to thank NASA / GSFC Dr. D, J. Cavalieri & Dr. J.C. Comiso, Former Professor Nishio of Chiba University, and Associate Professor Nagato of Tokai University, for their nice discussions as well as for providing synthetic aperture radar data. The author would like to extend my sincere gratitude to the Wakabayashi and other members of the Okhotsk Sea Ice Research Group. The author, also, would like to thank Professor Dr. Hiroshi Okumura and Professor Dr. Osamu Fukuda for their valuable discussions.

REFERENCES

[1] Kohei Arai, "Remote Sensing of Ice by the Earth Observation System (EOS)", Journal of the Remote Sensing Society of Japan, Vol.10, No.4, pp.124-128, 1990.

[2] Gloersen P., W. J. Campbell, D. J. Cavarieri, Global maps of sea ice cencentration, age and surface temperature derived from NIMBUS-7 SSMR:A case study, Oceanography from Space, Plenum Press, p.777,1981.

[3] Cavarieri D. J., A microwave technique for mapping thin ice, Journal of Geophysics Research, Vol. 99, pp. 12561-12572, 1994.

[4] Comiso J. C., SSM/I concentrations uisng the bootstrap algorithm, NASA RP 1380, pp.1-40, 1995.

[5] Comiso J. C., H. J. Zwally, Temperature corrected bootstrap algorithm, Proceedings of the IGARSS '97, pp.3835-3839,1997.

[6] Kohei Arai, Yasunori Terayama, Masao Matsumoto, Hiroki Fujiku, Kiyoshi Tsuchiya, Estimating Occupancy of In-Pixel Covering Class by Solving Inverse Problem, Saga University Bulletin, Vol.19, No.2, pp.9-15, 1991.

[7] Naofumi Ito, Sadao Fujimura, Area Ratio Estimation by Pixel Category Decomposition, Transactions of the Society of Instrument and Control Engineers Vol.23, No.8, pp.20-25, 1987.

[8] Masao Matsumoto, Hiroki Fuji, Kiyoshi Tsuchiya, Kohei Arai, Category decomposition based on maximum likelihood estimation, Journal of the Photographic Society of Japan, Vol.30, No.2, pp.25-34, 1991.

[9] K.Tachi, Kohei Arai and Y.Satoh, Advanced Microwave Scanning Radiometer -Requirements and Preliminary Design Study-, IEEE Trans. on Geoscience and Remote Sensing, Vol.27, No.2, pp.177-183, Jan.1989.

[10] Kohei Arai and Y.Terayama, Method for proportion estimation of mixed pixels by means of inversion problem solving, Proc.of the ISPRS Commission VII, WP-1-1, 1990.

[11] Kohei Arai, Inversion techniques for proportion estimation of mixels in high spatial resolution of satellite image, Proc.of the 29th COSPAR Congress, A.8-M-3.09, Solicited paper 1992.

[12] M.Matsumoto and Kohei Arai, Inversion for emissivity-temperature separation with ASTER data, Proc.of the 29th COSPAR Congress, A.4-M-4.05, 1992.

[13] Kohei Arai, E.Ishiyama and Y.Terayama, Method for ice concentration estimation with microwave scanning radiometer data by means of inversion, Proc.of the 30th COSPAR Congress, A3.1-032, 1993.

[14] Kohei Arai, E.Ishiyama and Y.Terayama, A method for ice concentration estimation with microwave radiometer data by means of inversion techniques, Advances in Space Research, Vol.16, No.10, pp.129-132, A31-32, Jul.1994.

[15] M.Moriyama and Kohei Arai, An Inversion for Emissivity-Temperature Separation with ASTER Data, Advances in Space Research, Vol.14, No.3, pp.67-70, Jul.1993.

[16] Kohei Arai, K. Teramoto and M. Matsumoto, Spatial Resolution Enhancement by Means of Inversion, Advances in Space Research, Vol.14, No.3, pp.71-74, Jul.1993.

[17] Kohei Arai, Inversion Techniques for Proportion Estimation of Mixels in High Spatial Resolution of Satellite Image Analysis, Advances in Space Research, Vol.14, No.3, pp.75-78, Jul.1993.

[18] Kohei Arai, Ice Concentration Estimation Based on Local Inversion, Proceedings of the IEEE Geoscience and Remote Sensing Symposium, 1997.

[19] Kohei Arai, Application of Inversion Theory for Image Analysis and Classification, Advances in Space Research, Vol.21, 3, 429-432, 1998.

[20] Kohei Arai and Masao Moriyama, SST Estimation Method with Linerized Inversion of Radiative Transfer Code for ADEOS/OCTS, Proc. of the COSPAR Congress, A0.1-0021, 1998.

[21] Kohei Arai and J.Sakakibara, Estimation of SST, wind speed and water vapor with microwave radiometer data based on simulated annealing, Advances in Space Research, 37, 12, 2202-2207, 2006.

[22] Kohei Arai, Nonlinear Optimization Based Sea Surface Temperature: SST Estimation Methods with Remote Sensing Satellite Based Microwave Scanning Radiometer: MSR Data, International Journal of Research and Reviews in Computer Science (IJRRCS) Vol. 3, No. 6, 1881-1886, December 2012, ISSN: 2079-2557, 2012.

[23] Kohei Arai, Simultaneous estimation of geophysical parameters with microwave radiometer data based on accelerated Simulated Annealing: SA, International Journal of Advanced Computer Science and Applications, 3, 7, 90-95, 2012.

[24] Kohei Arai, Sensitivity analysis for water vapor profile estimation with infrared sounder data based on inversion, International Journal of Advanced Computer Science and Applications, 3, 11, 65-70, 2012.

[25] Kohei Arai, Data fusion between microwave and thermal infrared radiometer data and its application to skin sea surface temperature, wind speed and salinity retrievals, International Journal of Advanced Computer Science and Applications, 4, 2, 239-244, 2013.

[26] Kohei Arai, Comparative Study of optimization Methods for Estimation of Sea Surface Temperature and Ocean Wind wit microwave Radiometer data, International Journal of Advanced Research on Artificial Intelligence, 5, 1, 1-6, 2016.

AUTHOR'S PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is a Science Council of Japan Special Member since 2012. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 55 books and published 620 journal papers as well as 450 conference papers. He received 66 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.html

The Effectiveness of Adopting e-Learning during COVID-19 at Hashemite University

Alaa Obeidat[1]
Hashemite University
Dept of basic Sciences
Zarqa, Jordan

Rana Obeidat[2]
Zarqa University
Faculty of Nursing
Zarqa, Jordan

Mohammed Al-Shalabi[3]
World Islamic Sciences and Education University
Amman, Jordan

***Abstract—e-Learning* is the utilization of the electronic technologies and the media to deliver the educational content to the learners, enabling them to interact actively with the content, the teachers, and their peers. Students' interaction can be either synchronous or asynchronous or a combination of both. One advantage of the e-learning is that learners can access the educational content at any place and time saving them effort, time, and cost. To deal with the unprecedented crisis of COVID-19 and the risk of virus transmission in the public, the vast majority of higher learning institutions globally were locked out and the delivery of the educational content moved from the traditional classroom teaching to the internet. The purpose of this study was to assess students' perceptions of the effectiveness of the e-learning during COVID-19 pandemic at the Hashemite University, Jordan. A total of 399 students completed the online survey of the study. Study results showed that students' overall evaluation of their e-learning experiences were generally positive. However, students reported that they faced problems in the e-learning experiences of which most were related to technical issues (e.g., lack of a viable internet network, lack of laptops, etc.). Microsoft Teams was the platform most preferred by students for e-learning and the majority of students accessed the educational content using smart phones. Only gender and student's academic specialty had significant associations with their perceptions of the effectiveness of the e-learning.**

Keywords—e-Learning; COVID19; classroom; ICT; Hashemite University; educational platform

I. Introduction

Information and Communication Technology (ICT) are vital to all aspects of everyday life in general and in education in particular. For its crucial role in creating an effective learning process and enhancing the role of learning, many educational institutions have adopted the use of ICT to continue with the process of educational communication. This transition from a traditional classroom to e- learning has also led to the emergence of new concepts within the world of education such as: e-learning, education through the internet, e-book, virtual university, e-library and other electronic media to allow the learner to learn according to personal preferences [1]. Learners can decide where and when they want to learn. They can prefer to learn without the commitment to attend classroom teaching at specific times. With the supply and accessibility of such modern technology in educational institutions, the integrated education using this technology has been designed and termed generally as e-learning.

e-Learning is not only an optimized use of both the human and material resources, but it also solves 'rare disciplines' problem. e-Learning makes it easy for the learner to communicate with any rare scientific discipline from anywhere in the world in no time. This is made possible as the training materials are prepared and made available by the educational institutions for those who need it. This is a low-cost solution to access and gain knowledge.

The deep-seated aspect to be considered before opting for e-learning is to check the feasibility for this adoption and investment. It would help to circumvent hurdles and troubles while transitioning traditional mode to information technology. Proper consideration is required to manage the resistance to change, aversion to transition and implement the change effectively [2]. To develop a successful e-learning strategy, such a system should be selected and developed appropriately that fulfills the requirements of education including constant updates to remain well-informed about the developments, observing standards and controls within the chosen education system to make sure that the level and development of the learner is under way and educational and non-educational goals are accomplished.

Information technology should not be mistaken for a goal. Rather, it is a way of communicating knowledge that is adopted and employed to achieve the purposes of education, and to enable the learner to combat the requirements of work-life with all its facets, which are becoming more dependent on information technologies and are evolving frequently.

A. Types of e-Learning

- Simultaneous Delivery

In this type, lecturer and students (learners) in different educational institutions such as universities, institutes and schools, communicate and interact directly at any real time. This mode can also be termed as synchronous and can be employed to conduct distance learning and some training courses.

- Asynchronous Delivery

In this type, the lecturer makes the course material available on videotapes, or transfers the contents via a computer or any other mean. While the student (the recipient) from the other side, receives the material at a later time suitable and feasible for him [3].

B. *Distance Learning Goals*

In line with the general aims of the education, distance learning aims to provide much more [4]. It promises:

1) Provision of educational opportunities for those who are deprived of these chances at all levels for many reasons that may include political, geographic, economic or social to name a few. Primarily, the distance education is to facilitate those ambitious learners who are striving to develop and educate themselves and to improve the educational, social and professional levels.

2) Creating apposite educational conditions to suit and cater to the needs of learners to aid continuous learning.

3) Providing the learners with the flexible schedule to fit to the conditions and circumstances for varied students such as housewives, farmers, industrialists and employees.

4) Redefining education, yet keeping it compatible with the provision of knowledge and the scientific and technological revolution in the present era. This new concept of education must enable the individuals to achieve competence through continuous education and self-learning. The idea of 'anytime, anywhere' learning is the need of the time, so that the students can learn at the moment they want it using the technology they prefer which is different from a traditional classroom and lecture hall environment.

5) Introducing new specialties that society needs, but traditional college systems do not allow and support to achieve.

6) Providing all citizens with cultural programs and knowledge. The benefits of information and communication technology are not just limited to students only, but all citizens can get benefits. It is possible because of the modern means of communication for example, television and satellites and using them to broadcast educational programs through them.

C. *Elements of Distance Learning*

To start with, distance learning requires the availability of the internet to facilitate communication with the student or learner present and responsible to track everything related to the educational material. Specially designed sites and portals following an appropriate mechanism can be used to achieve the purpose. These sites and portals should explain the material in an easy to understand manner so that they can be proved beneficial. Discussion forums, both direct and indirect, can also be made available to the teacher and the student. Lastly, teacher who has been assigned the responsibility to monitor and evaluate the student's performance must be available and must award marks to the student that he deserves.

D. *Distance Learning Methods*

There exist many methods of distance learning, and each of these methods target a specific stage of the educational interaction during the development of distance learning. Development in information and communication technology is taking place at a huge pace. This development and evolution is reflected in its educational uses that are expanding. New and more effective methods of distance learning are emerging [5] [6]. Among the most proven methods for distance learning include:

- Multimedia Style:

This method is based on written communication. Learners make use of the written material provided to them through audio and video recordings on CDs or the phone and radio broadcasting. Educational references, study guides and systematic books are printed and provided to the learners. Multimedia style can not only be adopted to support distance learning on its own but it also provides substantial support to other methods as well.

- Video Conferencing Method:

As it is stated in [7], this method is same as the learning happening in a traditional classroom. However, learners are geographically distant from their teachers and colleagues and connected via high-speed communication channels. Everyone can see and listen to the teacher. They can ask direct questions. They can actively participate in a discussion and can interact with the topic presented by the teacher. But to ensure everything goes well and as planned, the video conferencing method requires prior planning and preparation and takes longer time as compared to a traditional class lecture. It also demands scientific material and media. Teachers should be trained to attain student's attention and interest. All these are prerequisites to make effective use of this technology.

II. LITERATURE REVIEW

A. *e-Learning*

e-Learning is a new form to learn and communicate and it employs new computing technologies with high speed networks. It facilitates to develop skills, learn distantly and collaborate globally. These emerging technologies have evolved and transformed the learning and the processes and environments involved in learning. Author in [8] defines e-learning as the use of web-based technology tools along with the materials available on CDs, the internet, video and audiotapes and TV Broadcasts. Whatever tool it may use, main objective of e-learning is to provide a learning opportunity for individuals. [9] Looks for a single definition of e-learning which does not exist due to the different perspectives and constant evolution of this concept. e-Learning is found to be an amalgamation of various disciplines including data communication, computer science and pedagogy. So, the e-learning must be defined in the vast terms. Author describes e-learning as far more than just the technology and considers four main categories while defining it: "technology, delivery systems, communication, and educational paradigms".

B. *e-Learning during COVID-19*

COVID-19 pandemic led to the emergency closure of the educational institutions all around the world. To continue the learning process employing e-learning and online classes is a real challenge for governments and academic institutions. The author in [22] explains how lack of technical support, awareness, readiness, skills, resource materials and infrastructure pose challenges to the adoption of e-learning. Trust issues, resistance to accept the change and financial issues are also critical to the success of e-learning programs.

C. Effectiveness and Adoption of e-Learning among the Learners and Teachers

S. Bell, B. & E. Federman, J [10] explains the widespread use of ICT in recent and future years. Authors declare that comparing e-Learning's effectiveness with a traditional classroom is an old research topic. Focus should now be shifted towards making this mode of learning more effective and enhancing its features. Future research and advancements should modify the design model of e-learning system to overcome existing barriers and widespread the use and adoption of "online instruction" across cultural and geographical boundaries. Author in [11] examines the factors that impact the adoption of the e-learning programs and it is found that adoption is decided by some factors including: "relative advantages, trial ability, and academic specialization". Positive relative advantage increases the adoption rate for new systems and technologies. If the teachers are given a chance to test the tools and technology before implementation, they feel more confident and thus adoption rate improves. Further, the faculty from IT-oriented departments is found to be more welcoming to adopt new technologies. Study finds age and experience irrelevant to the adoption rate which is conflicting with many existing researches. Author in [12] identifies underlying factors that influence students' intentions to use new learning technologies involving collaborative features. The enhanced model declares peer influence, ease of use, compatibility with existing tools and practices significantly affect the adoption rate.

In higher-education, e-readiness is considered as a critical factor for the success of e-learning programs. Author in [13] discusses the e-readiness factors and improvement of those factors to make e-learning more effective. The study discovers skills and attitudes as the most crucial factors influencing e-learning readiness. Author in [14] acknowledges that in an online class, students may find it difficult to communicate with other students, faculty and administration. Even some students isolate themselves or somewhat feel isolated. They don't come forward and ask for any assistance on their own. Students face challenges in interaction and collaboration in an online classroom because of absence of fact-to-face and a non-verbal communication. Students don't feel themselves really connected. Talking about the advantages, while planning for online learning programs, curriculum can be designed keeping a particular student in mind. Study finds that to ensure connectivity and engagement in an online class, focus should be on building trust, showing presence, encouraging teamwork, and improving personal contact.

Declares e-learning as a new era in education which has revolutionized thoughts, enhanced existing educational and training models where the learner occupies a central and pivotal position [15]. Although ICT has its own disadvantages and limitations, but it emphasizes on learning that is independent, flexible, interactive and fits in with the time and pace of a learner. To adopt this mode in more effective way, infrastructure must be improved. Author in [16] finds e-learning truly effective when it incorporates and caters to the individual needs of the learner. In such an environment, online learners' exhibit enhanced capabilities and more organized thinking processes when they are given individualized instruction. Author in [17] studies the impact of various factors on the effectiveness of e-learning. Factors like the level of education, age, study program and gender are found to have significant influence on the success of e-learning. However, some factors are found to be irrelevant including race and marital status. It is indicated that learners enrolled in high level of education accept this new mode of learning easily and contribute towards its effectiveness. While considering the gender, female learners are found to be more patient and benefit more from the e-learning programs.

Discusses the factors that define and determine the success and effectiveness of e-learning [18]. Paper declares the use and user satisfaction of e-learning as closely linked. The user satisfaction is found to be related to quality of resource material and system, teacher attitude, evaluation and assessment methods, and learner's interaction with peers and the teacher. Use of e-learning is more effective when quality of collaboration and content, and above all, user satisfaction are improved. Technology should be employed to ensure collaborative environment. Similarly, the information quality also significantly improves the e-learning programs. Contents that are accessible, useful, easy to understand and authentic enhance the use and the user satisfaction. Assessment can be done in a variety of ways like using quizzes and tests. Similarly, overall success of the program can be made possible by improving these factors. e-Learning programs can be made more effective and successful if they take into consideration the individual requirements and interest of the learners. Author in [19] proposes that e-learning can only be made personalized with deliberate efforts to fit in to the needs, objectives, interests, and abilities of a learner. This paper recommends a programmed online tutoring system which adapts itself automatically to the interests and abilities of the learners. For the purpose, this personalized e-learning system records the knowledge, learning style, choices and objectives of an individual leaner. Author in [20] presents a six-dimensional integrated model for e-learning discovering factors affecting the user satisfaction. These dimensions include: "learners, instructors, courses, technology, design, and environment." Study concludes that uneasiness of learner with the electronic gadgets, teacher's mind-set towards e-learning, course flexibility and quality, navigation complexity, variety of ways to assess and evaluate the learner are the factors that must be focused to improve the e-learning programs.

Discusses about the failures of e-learning portals and initiatives in developing countries [21]. Reasons behind the failure are found to include lack of technical knowledge, not-easy-to-use systems and technologies, ineffective marketing, and insufficient training and support. Mainly, the study finds that failure is not because of technology or its availability. Rather, the improper use of technology causes the failure. e-Learning systems are failing to satisfy the users for not serving the purpose they were developed for. Paper also discusses possible way out to improve the system.

D. Study Purpose and Research Questions

This study aimed to assess students' perceptions of the effectiveness of the e-learning during COVID-19 pandemic at the Hashemite University, Jordan. The study was conducted to answer two research questions:

RQ 1*:* What is the level of effectiveness of using E- learning during the era of COVID-19 at the Hashemite University?

RQ 2**:** Are there statistically significant differences in the responses of the study sample individuals to the effectiveness of using e-learning during the COVID-19 era at the Hashemite University according to the variables (e.g., gender, student's academic year, specialization, place of residence, GPA, branch of education in the Tawjihi)?

III. Materials and Methods

A cross-sectional survey research design was used to answer the research questions of this study. A total of 399 students from different academic fields at the Hashemite University, Jordan were conveniently recruited and completed the study survey. An online survey (i.e. Google forms) was used to collect data from the study respondents. Students were invited to the study using different social media and platforms (e.g., Facebook, university website). The online survey was composed of three parts: a study cover letter, a demographic data sheet and a questionnaire about students' perceptions about e-learning. The study cover letter contained information about the purpose of the study, risks and benefits, assurance of keeping students' information confidential, and contact information of the researcher. Students' completion of the online survey was considered their consent to participate in the study. The questionnaire contained a total of 19 items which asked about students perceptions related to the impact of distance learning process, the amount of time each student spends on different platforms, any challenges students faced in the e-learning process, and their attitudes toward the potential continuation of the e-learning in the next 6 months if the COVID-19 pandemic did not come to an end. Students were asked to indicate their agreement with each of the 19 statements in the questionnaire using a Liker scale ranging from 1 (strongly disagree) to 4 (strongly agree). Data were analyzed using the Statistical Package for Social Sciences (SPSS). Descriptive (i.e. means, standard deviations, and frequencies) and inferential statistics (t test and one-way analysis of variance) at a significance level of .05 were calculated as appropriate.

IV. Results

A. Statistical Standard

Lecture quadrant is approved to correct the study tools, by giving each of its paragraphs one of the four degrees (strongly agree, agree, disagree, strongly disagree) and they are digitally represented (4, 3, 2, 1) respectively, and it has been done. To analyze the results, the following scale is adopted:

- 0.00 through 2.00 is low
- 2.01 through 3.00 is average
- 3.01 through 4.00 is high

The following formula is used to calculate the scale:

Upper limit for scale (4) - Lower limit for scale (1)/ Number of required categories (3)

(4-1) /3 = 1

Then add the answer (1.00) to the end of each category.

B. The Study Sample

The sample of the study is distributed (see Table I).

TABLE I. FREQUENCIES AND PERCENTAGES

Variables	Categories	Frequency	Percent
Student's school year	First	110	27.6
	Second	92	23.1
	Three	91	22.8
	Fourth	97	24.3
	Fifth	9	2.3
Specialization	Medicine	65	16.3
	Languages	53	13.3
	Medical and health sciences	40	10.0
	Engineering	11	2.8
	Sport	38	9.5
	College of Education with its specializations	23	5.8
	College of Sciences with its specializations	20	5.0
	Information technology with specializations	7	1.8
	Others	142	35.6
Sex	Male	137	34.3
	Female	262	65.7
Address	City	317	79.4
	Village	74	18.5
	Camp	8	2.0
GPA	Less than 2	7	1.8
	2- 2.5	70	17.5
	2.6-3	129	32.3
	3.1-4	193	48.4
Academic branch in Tawjihi	Literary	156	39.1
	Scientific	174	43.6
	Industrial	17	4.3
	Others	52	13.0
Total		399	100.0

C. First: Descriptive Analysis of the Study Tool

For the estimates of the study sample individuals on the effectiveness of using e-learning during the COVID-19 pandemic at the Hashemite University to indicate the level of their response to the items of the tool, the arithmetic averages and standard deviations are calculated.

See Table II, it shows the arithmetic averages between 1.87 and 3.14. Paragraph Number (10) which states, "There are problems that I face during distance learning, most of which are technical" comes in the first place with an arithmetic average of (3.14), and Paragraph Number (7), which states "Distance learning helped me to understand the material in a way more than the traditional method," appears last, with an arithmetic average of (1.87), and the arithmetic mean of the tool as a whole was (2.37).Thus, the majority of the students agreed that using e-learning programs during the era of COVID-19 is effective on average.

TABLE II. ARITHMETIC AVERAGES AND STANDARD DEVIATIONS ARRANGED IN DESCENDING ORDER ACCORDING TO AVERAGES

Rank	No.	Items	Mean	Std. Deviation	Level
1	10	There are problems that I encounter during distance learning, most of which are technical	3.14	.715	High
3..3.0 2	11	I can ask questions during the distance learning and get my answers	2.90	.733	Average
3	1	Ican access the electronic lectures easily	2.64	.850	Average
4	3	Provides technical support from the university in the event of a technical problem	2.63	.775	Average
5	12	Distance learning helped me enhance my self-learning ability	2.62	.897	Average
6	4	The electronic content displayed for the material is sufficient appropriate	2.54	.850	Average
7	16	Distance learning was a way to focus more on studying because it provided time to go to university	2.49	1.034	Average
8	6	Lectures include assignments and exercises that facilitate the understanding of the material	2.48	.789	Average
9	13	Use different assessment methods that suit me during distance learning	2.43	.820	Average
10	8	Distance learning has provided me with new skills	2.26	.889	Average
11	9	Distance education provided me with thinking and problem-solving skills	2.26	.895	Average
12	18	Distance learning is more flexible than traditional learning	2.21	.987	Average
13	2	Provides internet service with reasonable speed and without problems	2.19	.893	Average
13	5	The presentation of the electronic lecture is interesting and interesting	2.19	.860	Average
15	15	The level of interaction with the teacher was more than the traditional method	2.16	.927	Average
16	17	I prefer diversity using educational platforms instead of using only one platform	2.11	.935	Average
17	19	In general, I prefer to continue with distance learning	2.03	1.117	Average
18	14	I think remote exams were fair and measured all students at the same level	1.95	.990	low
19	7	Distance learning helped me to understand the subject more than the traditional method	1.87	.935	low
		The tool as a whole	2.37	.587	Average

D. Second: Questionnaire questions

1) From your point of view, which of the learning platforms is the best and least technical problem?

As shown in Fig. 1 and Table III that the students prefer to use MS Teams as medium of communication more than the other tools.

2) What is the method used to access the learning platform?

As shown in Fig. 2 and Table IV that the students prefer to smart phones to access to the different educational platforms and to download the courses materials.

E. Third: Finding Statistically Significant differences Attributable to Demographic Variables

Arithmetic averages and standard deviations are extracted for the responses of the study sample in the effectiveness of using e-learning during the era of COVID-19 in the Hashemite University according to variables (gender, student's school year, specialization, place of residence, grade point average, and academic branch in the guideline), and Table V shows that.

Table V shows significant statistical differences at the significance level (p = 0.05) due to the effect of gender on the responses of the study sample on the effectiveness of using e-learning during the COVID-19 pandemic at the Hashemite University, and the differences appear in favor of males.

Table VI shows a clear variation in the arithmetic averages and standard deviations of the study sample responses in the effectiveness of using e-learning during the COVID-19 era at the Hashemite University due to the different categories of variables (the student's school year, specialization, place of residence, GPA, branch of Tawjihi). To illustrate the significance of the statistical differences between the arithmetic averages, the five-way variance analysis is used according to (See Table VI).

It is clear from the Table VII that there exist no significant statistical differences at the level of significance (□ = 0.05) due to the variables: the student's school year, place of residence, grade point average, and academic branch in the Tawjihi.

It is also found that there are present significant statistical differences at the level of significance (□ = 0.05) due to the variable of study on the responses of the study sample in the effectiveness of using e-learning during the era of COVID-19 in the Hashemite University, and to show the significant statistical differences between the arithmetic averages, the dimensional comparisons are used in an oral manner as shown in the Table VII.

Table VIII it is apparent that there are considerable statistical differences (p= 0.05) between the specialty of languages and others on the one hand and the specialty of medicine on the other hand, and the differences come in favor of the medicine specialty.

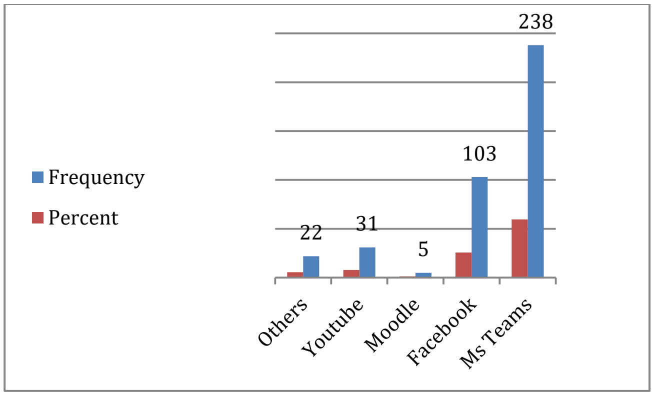


Fig. 1. Educational Platforms Preferences.

TABLE III. FREQUENCIES AND PERCENTAGES ACCORDING TO THE RESPONSE REGARDING PLATFORMS

Categories	**Frequency**	**Percent**
MS Teams	238	59.6
Facebook	103	25.8
Moodle	5	1.3
YouTube	31	7.8
Others	22	5.5
Total	399	100.0

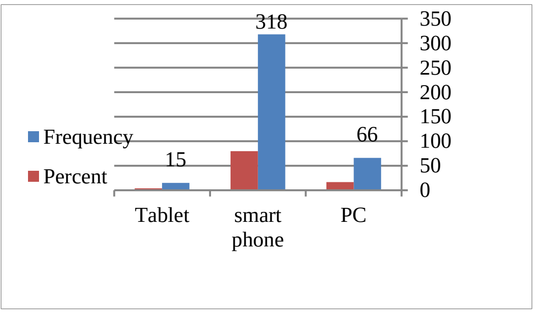


Fig. 2. Hardware Tool to Access the Platforms.

TABLE IV. FREQUENCIES AND PERCENTAGES ACCORDING TO THE RESPONSES REGARDING HARDWARE TOOLS

Categories	**Frequency**	**Percent**
PC	66	16.5
Smart Phone	318	79.7
Tablet	15	3.8
Total	399	100.0

TABLE V. AVERAGE, STANDARD DEVIATIONS AND T-TEST FOR SEX IMPACT

Variable	**Categories**	**N**	**Mean**	**Std. Deviation**	**t**	**Df**	**Sig**
Gender	Male	137	2.51	.636	3.499	397	.001*
	Female	262	2.30	.547			

TABLE VI. MATHEMATICAL AVERAGES AND STANDARD DEVIATION OF IMPACT (THE STUDENTS` SCHOOL YEAR, MAJOR, PLACE OF RESIDENCE, GPA, BRANCH OF TAWJIHI))

Variable	**Categories**	**N**	**Mean**	**Std. Deviation**
Student's school year	First	110	2.34	.524
	a second	92	2.46	.562
	Three	91	2.31	.616
	Fourth	97	2.37	.632
	Fifth	9	2.45	.759
Specialization	Medicine	65	2.68	.494
	Languages	53	2.26	.510
	Medical and health sciences	40	2.28	.509
	Engineering	11	2.19	.570
	Sport	38	2.51	.720
	College of Education with its specializations	23	2.34	.505
	College of Sciences with its specializations	20	2.47	.702
	Information technology with specializations	7	2.52	.939
	Others	142	2.27	.565
Address	City	317	2.39	.584
	Village	74	2.30	.608
	Camp	8	2.30	.535
GPA	Less than 2	7	2.65	.346
	2- 2.5	70	2.48	.732
	2.6-3	129	2.34	.573
	3.1-4	193	2.35	.539
Academic branch in Tawjihi	Literary	156	2.23	.537
	Scientific	174	2.48	.556
	Industrial	17	2.57	.739
	Others	52	2.38	.694

TABLE VII. FIVE-YEAR VARIANCE ANALYSIS OF THE IMPACT OF (THE STUDENTS` SCHOOL YEAR, MAJOR, PLACE OF RESIDENCE, GPA, BRANCH OF TAWJIHI)

	Type III Sum of Squares	**df**	**Mean Square**	**F**	**Sig.**
Student's school year	1.446	4	.362	1.130	.342
Specialization	7.882	8	.985	3.081	.002*
Address	.439	2	.220	.687	.504
GPA	2.040	3	.680	2.126	.096
Academic branch in Tawjihi	1.318	3	.439	1.374	.250
The error	120.894	378	.320		
Kidney	137.134	398			

TABLE VIII. DIMENSIONAL COMPARISONS IN A HEALING MANNER TO THE EFFECT OF SPECIALIZATION

Categories	Medicine	Languages	Medical and Health Sciences	Engineering	Sport	College of Education: Specializations	College of Sciences: Specializations	Information technology: Specializations	Others
Medicine	1								
Languages	.42*	1							
Medical and health Sciences	.40	.02-	1						
Engineering	.49	.06	.09	1					
Sport	.17	.25-	.23-	.32-	1				
College of Education: Specializations	.34	.08-	.06-	.15-	.17	1			
College of Sciences: Specializations	.22	.21-	.19-	.27-	.04	.12-	1		
Information technology: Specializations	.16	.26-	.24-	.33-	.01-	.18-	.05-	1	
Others	.42*	.00	.02	.07-	.24	.08	.20	.25	1

V. DISCUSSION AND CONCLUSION

Information and communication technology is a new tool in education and learning especially in less developed countries like Jordan. Because of the unprecedented crisis and the threat to public health imposed the COVID-19 virus and thus the national lockdowns, the vast majority of higher educational institutions across the globe had to move to the e-learning to promote the continuity of the teaching-learning process. e-Learning programs are not new to the teaching learning process as there are both undergraduate and graduate programs that are completely delivered electronically (e.g. distance learning, online learning, or a combination of both) in many countries across the world especially in developed countries. Thus, the transition to e-learning during the COVID-19 pandemic might not be difficult for these countries and thus could be better adopted and received by both teachers and students in their countries. On the other hand, as this study found, in other countries especially developing ones like Jordan where the technological infrastructure might not be maturely developed, the transition to e-learning might not be well received and adopted. Despite that student respondents involved in this study were overall satisfied with the e-learning they also reported a plethora of difficulties that faced during the e-learning process of which most were related to technical problems. Those technical problems might have been caused by the immaturity of electronic technologies in Jordan compared to more developed countries. For instance, a viable internetwork might not be available in many rural and remote areas in the country. Also, both teachers and students in Jordan are not accustomed to e-learning and distance learning which could be another factor affecting the effectiveness of the e-learning as reported by student respondents in this study. Thus, it is essential that policy makers (e.g., Hashemite University administration)in the country take challenges reported by student respondents in this study into consideration when planning for future interventions to enhance technology readiness for e-learning programs, improve students' confidence in adopting and transitioning to this new mode, and thus to make e-learning more effective.

Results of this study revealed that only students' gender and academic specialty were associated with their evaluation of the effectiveness of the e-Learning. Student's school year, place of residence, grade point average, and academic branch in the Tawjihi were found to have no significant associations with the effectiveness of e-Learning. Students specializing in medicine appear to be more welcoming of e-Learning. Medical students have accepted the e-learning better than students in other specialties and thus reported that e-learning contributed positively to their learning experiences and learning outcomes. This finding could be possibly attributed to differences in performance expectations (e.g., more integration of cognitive activities) and learning climate/social environment between medical school and other schools in the university. In addition, male student respondents have accepted e-learning more smoothly and with patience and evaluated the experience more positively compared to female students. This result could be related to less exposure and the use of technology tools among female students compared to male ones. Further, female students could have suffered more anxiety because of the sudden change from the traditional class room teaching to e-learning which might have thus affected their performance and their satisfaction with the experience as a whole.

VI. FUTURE WORK

As future work, we will conduct a training sessions for the students to teach them the best techniques on the use of the educational platforms and how to overcome difficulties.

ACKNOWLEDGMENTS

The authors would like to show their gratitude towards Hashemite University who shared its pearls of wisdom with the authors during different questionnaires and research. The authors are also immensely grateful to the students on their comments and answers to different questionnaires that greatly improved the manuscript.

REFERENCES

[1] Noor-ul-Amin, S. (2013). An effective use of ICT for education and learning by drawing on worldwide knowledge, research and experience: ICT as a change agent for education (A Literature review). Scholarly Journal of Education.2 (4): 38-45.

[2] Ali, M., Zhou, L. &Ieromonachou, P. (2016). User resistance in IT: A literature review. International Journal of Information Management. 36(1): 35-43.

[3] Hrastinski, S. (2008). A study of asynchronous and synchronous e-Learning methods discovered that each supports different purposes. EDUCAUSE Quarterly: 31(4).

[4] Santana de Oliveira, M.M., Torres Penedo, A.S. & Pereira, V. (2018). Distance education: Advantages and disadvantages of the point of view of education and society. Dialogia. 29: 139-152.

[5] Zawacki-Richter, O., Bäcker, E. M., & Vogt, S. (2009). Review of distance education research (2000 to 2008): analysis of research areas, methods, and authorship patterns. International Review of Research in Open and Distance Learning. 10(6): 21–45.

[6] Burns, M. (2011). Distance Education for Teacher Training: Modes, Models and Methods, Washington, DC: Education Development Center.

[7] Motamedi, V. (2001). A critical look at the use of videoconferencing in United States distance education. Education. 122: 386–394.

[8] L. Moore, J., Dickson-Deane, C. &Galyen, K. (2010). E-Learning, online learning, and distance learning environments: Are they the same? The Internet and Higher Education. 14: 129-135.

[9] Sangrà, A., Vlachopoulos, D. & Cabrera, N. (2012). Building an Inclusive Definition of E-Learning: An Approach to the Conceptual Framework. International Review of Research in Open and Distributed Learning, 13:145-159.

[10] S. Bell, B. & E. Federman, J. (2013). E-Learning in Postsecondary Education. Future of Children. 23 (1): 165-185.

[11] Mohamad Hsbollah, H. & Md. Idris. K. (2009). E-Learning adoption: the role of relative advantages, trialability and academic specialization. Campus-Wide Information Systems. 26(1): 55-70.

[12] Cheung, R. & Vogel, D. (2013). Predicting user acceptance of Collaborative technologies: An extension of the technology acceptance model for e-Learning. Computers and Education. 63: 160-175.

[13] Rohayani, A., &Kurniabudi, Sh. (2015). A literature review: Readiness factors to measuring e-Learning readiness in higher education.Procedia-Computer Science.59: 230-234.

[14] Swanson, A., Davis, B., Parks, O., Atkinson, S., Forde, B. & Choi, K. (2015). Student engagement, e-connectivity, and creating relationships in the online classroom: emerging themes. International Journal of Instructional Technology and Distance Learning, 12: 66-73.

[15] Talebiana, S., MovahedMohammadia, H. &Rezvanfara, A. (2014). Information and communication technology (ICT) in higher education: advantages, disadvantages, conveniences and limitations of applying e-Learning to agricultural students in Iran. Procedia - Social and Behavioral Sciences, 152: 300-305.

[16] Xu1, D., W. Huang2, W. Wang, H. &Heales, J. (2014). Enhancing e-Learning Effectiveness Using an Intelligent Agent-Supported Personalized Virtual Learning Environment: An Empirical Investigation. Information and Management, 51: 430-440.

[17] Md. Islam, A., Asliza Abdul Rahim, N., Chee Liang, T. &Momtaz, H. (2011). Effect of Demographic Factors on E-Learning Effectiveness in a Higher Learning Institution in Malaysia. International Education Studies, 4: 112-121.

[18] AudyeCidral, W., Oliveira, T., Di Felice, M. & Aparicio, M. (2018). E-Learning success determinants: Brazilian empirical study. Computers & Education. 122: 273-290.

[19] Klasnja-Milicevic, A., Vesin, B., Ivanovic, M. &Budimac, Z. (2011). E-Learning personalization based on hybrid recommendation strategy and learning style identification. Computers & Education. 56: 885-899.

[20] Chen Sun, P., J. Tsai, R., Finger, G., Yang Chen, Y. & Yeh, D. (2008). What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction. Computer and education.50: 1183-1202.

[21] Ssekakubo, G., Suleman, H. & Marsden, G. (2011). Issues of Adoption: Have E-Learning Management Systems Fulfilled their Potential in Developing Countries?In Proceedings of the South African Institute of Computer Scientists and Information Technologists Conference on Knowledge, Innovation and Leadership in a Diverse, Multidisciplinary Environment, Cape Town, South Africa.: ACM New York, NY, USA, 231-238.

[22] Amin Almaiah, M., Al-Khasawneh, A. & Al Thunibat, A. (2020). Exploring the critical challenges and factors influencing the e-Learning system usage during COVID-19 pandemic. Education and Information Technologies. 22 : 1–20.

Recovering UML2 Sequence Diagrams from Execution Traces

EL Mahi BOUZIANE[1], Abdeslam JAKIMI[3]
Software Engineering and Information Systems Team
Faculty of Sciences and Technics, My Ismail University
Errachidia, Morocco

Chafik BAIDADA[2]
Laboratory of Information Technologies
ENSA, Chouaib Doukkali University
El Jadida, Morocco

***Abstract*—Reverse engineering is a proven and efficient technique for automatically generating UML2 models from object-oriented legacy systems with missing or obsolete documentation. To perform reverse engineering, two techniques are used: dynamic and static analysis. Dynamic analysis refers to collecting information when the system is running while static analysis corresponds to inspecting the source code. Dynamic analysis is preferred than static one in order to extract dynamic models that represents the behavior of a systems because of polymorphism and dynamic binding. In this paper, we present new different methodology that use Colored Petri Nets (CPNs) to recover UML2 Sequence Diagram (SD). First, it generates execution traces corresponding to the different scenarios representing the system behavior. Then, CPNs are used to model and analyze these execution traces to extract UML2 sequence diagram. Our case study illustrates the process of our approach and show that sequence diagram can be extracted with a good accuracy.**

***Keywords*—Execution traces; Reverse engineering; UML2; Sequence Diagram; Colored Petri Nets**

I. Introduction

Today object-oriented systems, are becoming increasingly larger and more complex. This increases the cost of their development and maintenance. According to [1], the cost of software maintenance represents 50% to 75% of the total cost. Despite the progress made in software engineering and development methods, several legacy systems still suffer from many problems such as unavailability of developers, obsolete development methods used to code the software, outdated documentation and non-compliance with the design when coding the software. In the software lifecycle, understanding its architecture and behavior is the main task in the maintenance phases. It is a tedious and time-consuming task that requires the mobilization of a large number of human resources. As mentioned in [2], up to 60% of maintenance time is spent on understanding the software. Therefore, it is important to develop techniques to obtain an abstract representation that facilitate the understanding of these systems.

A proven and effective technique to face this problem is reverse engineering of UML2 models. It can be defined as a process of analyzing the source code of systems and representing it in models with a higher level of abstraction. Reverse engineering is mostly used to extract high level abstraction models or semantics from the source code [3]. Reverse engineering is used to help understanding existing systems. The IEEE-1219 [4] standard considers reverse engineering as a technological solution to deal with legacy system. For the object-oriented software, the most used modeling language is UML (Unified Modeling Language) [5]. Dynamic models are as important as static models because they allow to understand the behavior of the system. One of the major UML dynamic model is SD. Indeed, it allows to represent complex interactions between its objects [6]. As described in [7], dynamic analysis allows to remove the ambiguity of message sending when inheritance, delegation, polymorphism, dynamic links, reflection are used intensively. For this, we will give more importance to this type of analysis.

This paper draws on our previous work [8, 9] to propose a new, more coherent and precise approach for reverse engineering the UML2 SD. This new approach allows to extract the conditions for combined fragment operator alt, opt and loop. For this purpose, improvements in the generation of execution traces and modeling with CPNs have been made. Indeed, the CPNs used have a smaller size and are more coherent. However, this approach does not currently apply to multi-threaded systems.

The remainder of this paper is organized as follows. Section II includes related works. Section III introduces a background in reverse engineering of UML2 SDs using CPNs. Section IV outlines the proposed approach. Section V presents a case study. Finally, Section VI concludes and points out some of our future works.

II. Related Work

Reverse engineering is defined as “the process of identifying and analysis of software’s system components, their interrelationships, and the representation of their entities at a higher level of abstraction” [10]. Reverse engineering aims to discover the technological principles of a system through the analysis of its structure and behavior.

In the literature, depending on the type of analysis used, there are two main categories in existing approaches: static and dynamic. Static analysis consists in performing the analysis of the source code or the binaries to generate UML dynamic diagrams. This is done without running the system. There are several approaches that perform reverse engineering through static analysis [11, 12, 13, and 14]. One of the main works based on static analysis is [14]. In this work, the authors present an algorithm that builds UML 2.0 sequence diagrams from Java code, using control flow graphs. They create an algorithm that transform this graph on SD. This is done on three phases. First, the algorithm associates subgraphs with sequence diagram fragments operator alt, opt, break, and loop. After that, a series of transformations are applied to the obtained SD to in order to make it more understandable.

Dynamic analysis consists of running the program to obtain the necessary information in the form of an execution trace for the creation of sequence diagrams. These traces represent the values of the program variables, the state of the execution stack, the occurrences of objects created, the signatures of the methods called, the information about threads or any other execution information considered useful. As a result, objects under execution can be observed. There are several works that use dynamic analysis. In [15], an approach to extract SD from dynamic information of object-oriented programs is presented. In order to reduce repetitions in the execution trace, four rules are used to optimize the size of the execution traces by detecting similarity between sub-trees and replace merging them. The author in [16] propose an approach that allows extracting UML High Level Sequence Diagrams (HLSD) from java code by constructing control flow graphs. They proposed a method for switching between the general control flow graph (FCG) and UML sequence diagrams. The combination process is done by analyzing the different states of the system. In [17], it is proposed an approach based on dynamic analysis using Labeled Transition System (LTS). These LTSs are used for modeling execution traces in order to facilitate there analyzing. For each trace a corresponding LTS is generated. After that come the step to merge these LTSs in order to have one LTS modeling the behavior of the system. Finally, an HLSD is generated from the obtained LTS using regular expressions.

The approaches listed before were able to extract SDs that represent the system behavior. However, the diagrams obtained are incomplete and suffer from several problems. These problems include information filtering problems. As listed in the catalog of abstractions and filtering in the context of reverse engineering of sequence diagrams [18]. In addition, these approaches fail to extract the conditions in the combined fragment operators like loop, opt and alt.

III. Background of the Approach

In this section, we first explain what a sequence diagram in UML 2.x is. Second, we give some definitions regarding execution traces and how they are obtained. Finally, we introduce CPN and how we used it to represent an HLSD.

A. UML2 Sequence Diagrams

The SD is a form of behavioral diagram that allows to specify in a chronological way the interactions that exist between a group of objects from the temporal point of view. It has been significantly changed in UML 2.0 [5]. Indeed, the sequence diagram in UML2 is considered as partially ordered collections of events, which introduces new concepts such as combined fragment, parallelism and a synchronism and allows the definition of more complex behaviors.

An HLSD is obtained by combining Basic Sequence Diagram (BSD) using interaction operators. The most commonly used combined fragment operators in the UML2 sequence diagram are *seq* to express sequence, *opt* for optional, *alt* for alternatives and *loop* for iterative actions.

B. Execution Traces

Dynamic analysis, starts by generating traces. These traces are then analyzed to extract a HLSD. In our approach, for each scenario a trace execution is generated. In what follows, we introduce a set of definitions that are necessary to understand the approach.

Def. 1: A trace line is a method invocation or a control structure.

Def. 2: A method invocation is a triplet T1=<Sender, Message, Receiver> where:

Sender is the caller object, expressed in the form package:class:object.

Message is the invoked method of the receiver object, expressed in the form methodName (par1, par2, …).

Receiver is the called object, expressed in the form package:class:object.

Def. 3: A control structure is a triplet T2 = <controlType, status, condition> where:

ControlType has one of the following values: IF, ELSE, SWITCH, CASE or DEFAULT.

Status expresses the start or the end of the control structure.

Condition (optional) is the condition expression associated with IF, CASE, FOR, or WHILE.

Def. 4: (Equivalence between method invocations): The method invocations l1 = <s1, m1, r1> and l2 = <s2, m2, r2> are equivalent if and only if:

- The objects s1 and s2 (respectively, r1 and r2) are equivalent if they are instances of the same class.
- The messages m1 and m2 concern the same method and have the same arguments.

Def. 5: An execution trace is a set of trace lines.

An example of execution traces in the format described before are shown in Table 1. For each Scenario correspond a trace (ex: Trace1 refers to Scenario1). The trace Trace1 is composed of lines from L0 to L5. Lines L6 to L10 belongs to Trace2. Pack1 represent the packages to which classes A and B belong. m1() to m5() refers to the methods invocation (messages) of objects a, b, c, and d.

C. CPN

Petri nets is a formal modeling language used to represent the dynamic behavior of different systems (computer, industrial, telecommunications ...) [19]. It was first introduced in 1962 by the German mathematician and computer scientist Carl Adam. CPN is an extension of Petri nets. CPN is an extension of petri nets. This extension considerably reduces the size of the network when extending the modeling with Net Petri. It allows the distinction between places by attaching a color to them.

TABLE I. AN EXAMPLE OF TRACES

Trace1 L6. Pack1:B:b|m3()|Pack1:A:a L7. WHILE |BEGIN |condition1 L8. Pack1:B:b|m4()|Pack1:A:a L9. Pack1:B:b|m5()|Pack1:A:a L10. WHILE|END L6. Pack1:B:b|m3()|Pack1:A:a	**Scenario1**
Trace2 *L0. IF* | BEGIN | condition2 L1. Pack1:A:a |m1()|Pack1:B:b L2. ELSE | BEGIN L3. Pack2:C:c |m2()|Pack2:D:d L4. ELSE| END L5. IF | END L1. Pack1:A:a |m1()|Pack1:B:b	**Scenario2**

A Petri Net block is a subnet of the Petri Net with one initial place and one final place. Those places refer respectively to the precondition and the post-condition of the subnet. In [20], CPN is used to integrate scenarios represented in the form of SDs. They use four combined fragment operators (conditional, sequential, iterative and concurrent) to combine scenarios. CPNs are suitable for our approach as they can be transformed easily into an HLSD (see Fig. 1).

Transitions represent BSD or the operator such as seq, opt or control type as defined in Def 3. Places can represent a state of the system or the beginning or the end of the operator alt and loop. Colors are used to distinguish between traces.

Fig. 1 shows how a CPN can be transformed easily into a HLSD and vice-versa. Pi and Pf represent respectively the initial and the final place of the Petri Net block. The first transition a | m1() |b corresponds to the first BSD. In this BSD the object a of class A call to the object b of the class B with the message m1(). The place LOOP | BEGIN represents the state before the start of the operator loop that leads to two transitions. The transition IF | BEGIN| C1 allows entering in the loop statement. This is done when the condition C1 of loop is equal to the value true.

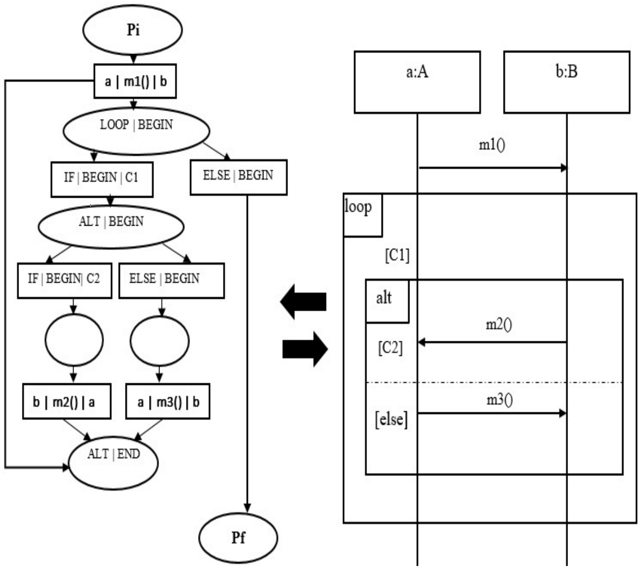


Fig. 1. A HLSD Mapped onto CPN with Operators Loop and Alt.

After that, comes the place ALT | BEGIN which represents the state that also leads to two transitions. The first transition labeled IF | BEGIN| C2 refers to the case when the condition of *alt* is satisfied and leads to the transition b | m2() | a. This transition describes that the message m2() is sent by the object a of the class A to the object b of class B. The transition ELSE | BEGIN| C2 represents the second transition of *alt* and consequently occurs when its condition is not verified. The transition a| m3() | b refers to the BSD which describes that the object a calls the object b using the message m3(). The second transition of *loop* is ELSE | BEGIN refers to its end and thus occurs when its condition is not verified.

In this section, we have presented and explained the most important concepts about SD, trace and CPN. This is the background on which our approach is based.

IV. PROPOSED APPROACH

Our objective is to extract SD from execution traces for an object oriented system using CPNs.

In this section, we present our approach for reverse engineering of the HLSD. As illustrated in Fig. 2, the approach is divided in four main steps. First, the step trace collection. Second, the trace filtering step. Third the step of trace merging. Finally, the step of HLSD extraction. In the next subsections, each step is described in details.

A. Traces Collection

In order to generate an accurate HLSD, our approach use dynamic analysis technic. In [6], it's described that dynamic analysis is more efficient than static one in the context of the reverse engineering of UML dynamic models such as SDs. This analysis is based on analyzing traces execution. These traces can be generated using several technics [2]. These technics includes the instrumentation of the source code, bytes code or the use of a customized debugger. In our approach we use the byte code instrumentation.

We choose to use AspectJ [21] as trace collection tools. This tool concerns java software systems. It allows to report all information created during the execution of the program. This includes methods invocations, occurrence of objects, sending and receiving messages between objects, loops and conditions.

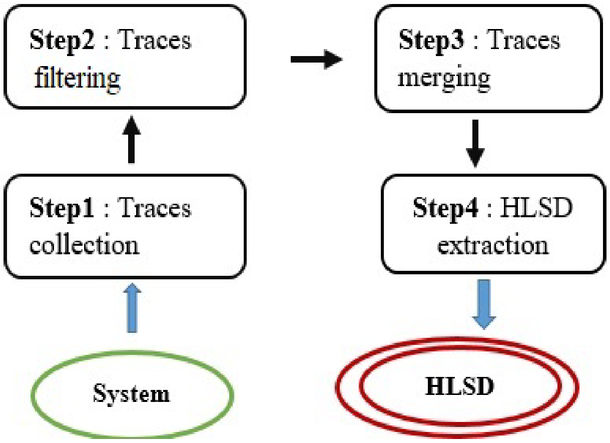


Fig. 2. Overview of our Approach.

The behavior of the system is highly dependent on the input data entered by the user. Therefore, it is necessary to identify the majority of input variable values in order to specify all system behavior. This can be done with the help of a system functional expert. This can be done with the help of a system functional expert. After running the system with different input data values, execution traces are generated, each corresponding to a given scenario. Since execution trace formats differ from one tool to another, we have developed a tool that will allow to standardize the format of execution traces. This format is defined in the definitions 1, 2, and 3. The objective of this tool is to format the trace execution to facilitate the processing of merging traces.

B. *Trace Filtering*

The generated execution traces contain a lot of information about all classes composing the system. For example, these classes can be divided into three types: data access classes, business classes and presentation classes. The business classes are the classes that describe the behavior of the business logic of the system. Our objective in this step is to concentrate on traces lines that describe this behavior and ignore other traces lines. This is the objective of the trace filtering step. We have developed an algorithm that allows us to delete execution traces which belong to data access or presentation classes.

C. *Traces Merging*

As mentioned before, one execution of the system doesn't allow an accurate description of all the system behavior. Therefore, the system must be run several times and thus generate different traces. The challenge is to be able to merge these different traces to identify the behavior of the system as a whole. In [22], several well-defined merging techniques were listed.

For merging execution traces, we choose to use CPNs. The process is done in two successive sub-steps: first CPN initialization, then CPN merging.

1) CPN Initialization: In this sub-step, a basic CPN is generated for each execution trace. All the trace lines are transformed into transitions in CPNs except those which refers to the start or the end of iterative control structure like LOOP | START and LOOP | END. These line traces are transformed into places. This reduces the size of CPN and makes it more consistent. A same color is assigned to all places that belong to the same execution trace. We use these colors to differentiate between scenarios. Each color refers to specific scenario. This allows subdividing an HLSD into less complex HLSDs to facilitate understanding the behavior of the system.

2) CPN Merging: In this sub-step, all the CPNs corresponding to execution traces are merged to obtain a single CPN. The algorithm kBehavior [23] is used for this reason. This algorithm has points in common with the known Ktail algorithm. [24]. These algorithms are used to construct finite-state automaton (FSA) that abstracts execution traces. The algorithms iteratively merge the equivalent states in order to generalize the resulting FSA. kBehavior can reuse already learned path to adapt it in the FSA with newly generated traces. This is not the case for Ktail. In our approach, we have developed an algorithm called adapted kBehavior which is totally inspired by kBehavior. It's a new version adapted to deal with CPNs. Adapted kBehavior does not need to pre-process the new traces it receives. However, it needs to explore the already generated CPN when it tries to learn again. When a new sequence of transition and places needs to be added to the CPN, it must be ensured that this sequence is not already present in the CPN. Since the generated CPN is generally not non-deterministic, the path of the CPN is quite inexpensive and the additional cost generated by this method remains reasonable.

To make the CPN more coherent, a final transformation is carried out. This transformation concerns the processing of an iterative behavior. This processing includes adding two test transitions after the place LOOP | BEGIN | CONDITION. The first transition labeled IF | BEGIN | CONDITION is executed when the condition of Loop is satisfied. The second transition labeled LOOP | END is executed in the other case. This transition leads to the place labeled LOOP | END and consequently indicating the end of Loop. The output place of the last transition inside *loop* does not refer anymore its end but to its beginning. The labeling of this place is changed by removing the indication of its condition in order to avoid redundancy as illustrated in Fig. 3.

D. *HLSD Extraction*

In this step, we can easily build an HLSD by mapping the resulting CPN using the following transformation rules.

- **Rule 1:** all names of objects in the CPN are transformed into lifelines in SD.
- **Rule 2:** a transition T1 with the method invocation 0:a:B | m1 ()| b:B is transformed into a BSD where object a:A sends message m1() to object b:B
- **Rule 3:** A Place P1 that contains the operator ALT | BEGIN **or** OPT |BEGIN **or** LOOP | BEGIN refers respectively to BSD with the operators alt, opt and loop.
- **Rule 4:** the CPN paths coming after the place ALT | BEGIN and ending on the transition ALT | END are transformed into combined fragments with the operators ALT.
- **Rule 5:** the CPN paths coming after the place OPT | BEGIN and ending on the transition OPT | END are transformed into combined fragments with the operators OPT.
- **Rule 6:** The cyclic CPN paths coming after the transition IF | BEGIN | CONDITION which comes after the place LOOP | BEGIN is transformed into combined fragments with the operator loop.
- **Rule 7:** The CPN paths coming after the transition ELSE | BEGIN | CONDITION which comes after the place LOOP | BEGIN is transformed into BSD after the fragment corresponding to the operator loop.

The rules listed below can be combined to map a CPN to an HLSD representing system behavior (Fig. 2).

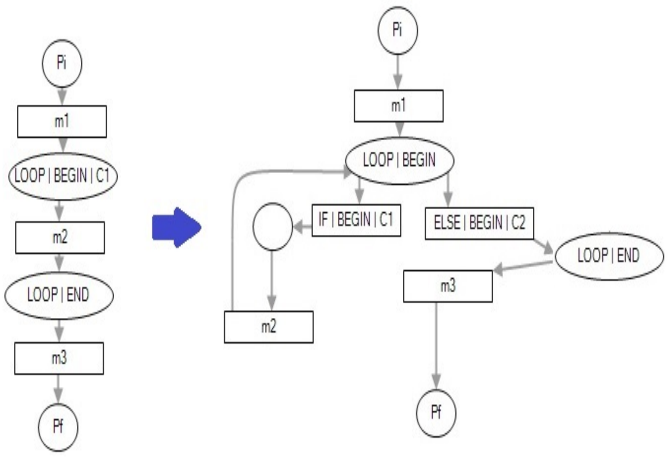


Fig. 3. CPN Corresponding to Scenario1.

V. Case Study

To test and illustrate the different steps of our approach, we have developed an application called Sales. It allows vendors to create sales of several articles. It gives the possibility to print an invoice, delivery or a payslip. All these operations are saved in a database. The application developed in Java provides different types of behavior (iterative, optional, sequential and alternative) which are the objective of our case study. The application has a layered architecture with two layers: business logic and data access layer and therefore is structured in two packages BLL (for business logic) and DAL (for data access layer). The BLL package contains six business classes (see Listing 1): Vendor, Sale, Calculation, Invoice, Payslip, and Delivery. The DAL package is composed of the following classes (see Listing 2): VendorDAL, SaleDAL, InvoiceDAL, PayslipDAL, and DeliveryDAL.

To create sales, the vendor makes an order to start a new sale. The vendor can add articles repetitively and calculate the total amount of the sales (repetitive behavior). When the vendor completes the sale, he chooses to print a delivery slip or an invoice in order to be signed (alternative behavior). Finally, if the customer wants it, a pay slip must be printed (optional behavior).

Listing1 and listing2 shows the source code of some classes of the application. Listing1 refers to the BLL package and Listing2 correspond to the DAL package.

Listing 1

```
package BLL;
import DAL.*;
class Vendor {
 public  Vendor(){
 }
  public static void signInvoice(){
  System.out.println("Invoice signed");
aVendorDAL.saveUpdate(this);
 }
 public static void signDelivery(){
  System.out.println("Delivery signed");
aVendorDAL.saveUpdate(this);
 }
 public static void signPayslip(){
  System.out.println("Payslip signed");
aVendorDAL.saveUpdate(this);
 }
}
class Sale {
 public void newSales(Vendor V, int nbr_article, boolean
isInvoice,   Boolean isPayslip){
  float oldValue=0;
  System.out.println("New sale created");
  for(int i=1;i<=nbr_article;i++){
   float newValue=addArticle();
   Calculation calcul= new Calculation();
   oldValue=calcul.calculAmount(newValue, oldValue);
  }
  if(isInvoice){
  Invoice invoice= new Invoice();
  invoice.waitMessage();
  invoice.start();
 }
 else {
 Delivery delivery= new Delivery();
 delivery.getDelivery();
 }
 if(isPayslip){
  Payslip payslip=new Payslip();
  payslip.getPayslip();
 }
 }
 public  float addArticle(){
  SalesDAL saleDAL =new SalesDAL();
  saleDAL.saveUpdate(this);
  System.out.println("New article added");
  return 1000;
 }
}
class Invoice extends Thread {
  public void preparingInvoice(){
   System.out.println("Preparing Invoice ");
 }
  public void waitMessage(){
   System.out.println("waiting for Invoice ");
 }
 public void getInvoice(){
 System.out.println("Invoice printed");
 Vendor.signInvoice();InvoiceDAL Inv1= new
InvoiceDAL();
 Inv1.saveUpdate(this);
 }
 @Override
   public void run() {
    try {
          preparingInvoice();
          Thread.sleep(1000);
          getInvoice();
    }
    catch (InterruptedException ex) { }
 }
}
class Payslip {
 public void getPayslip(){
  System.out.println("Payslip printed");
  Vendor.signPayslip();
  PayslipDAL Pay1 = new PayslipDAL();
  Pay1.saveUpdate(this);
 }
}
```

Listing 2

```
package DAL;
import BLL.*;

class DeliveryDAL  {
public  void saveUpdate (Delivery d){
   System.out.println
("Delivery updated in database  ");
     /*
...sending sql query to database using jdbc */
 }
}

class SaleDAL {
// if new one a sale is created and updated
// else the existing sale is updated
 public  void saveUpdate (Sales s){
   System.out.println
("sales updated in database  ");
     /*
 ... sending sql query to database using jdbc */
 }

public class PayslipDAL {
 public  void saveUpdate (Payslip s){
 System.out.println
("Payslip updated in database  ");
     /*
... sending sql query to database using jdbc */
 }
}

class VendorDAL {
 public   void saveUpdate (Vendor s){
   System.out.println("Vendor updated in database  ");
   /*
  ... sending sql query to database using jdb */
 }
}

```

A. *Trace Collection*

At this stage, the generated execution traces that represent the behavior of the systems are organized in text files according to the format proposed by our approach. To do that, we use a java program that we have developed. It takes as input the traces generated after instrumentation with AspectJ. Then, it adapts them according to the adequate format as it shown in Table II, Table III and Table IV.

B. *Trace Filtering*

As input for this step, we have formatted traces that each correspond to a given scenario. These traces include calls of all objects that belong to packages BLL and DAL. In this case study, we consider that the main behavior of our application is illustrated in the business logic layer. So, we will ignore all trace lines that refers to the data access layer (trace lines with red color). For that, the algorithm will delete all lines traces that includes the package: DAL. Therefore, the line traces in red in Table II will be deleted. The final traces will contain only traces that include the BLL package.

C. *Trace Merging*

This step consist in generating for every filtered trace a corresponding CPN (Fig. 4, 5 and 6). These CPNs include as transitions the events generated by the system like the invocation of methods and performing tests. The places of CPN contains only the stars and the end of structure controls: LOOP | BEGIN |, LOOP | END, ALT | BEGIN |, ALT | END. To do this, we have developed an algorithm that transforms every trace into a CPN.

First, our algorithm creates the initial place that represents the start of the CPN. Then, if it finds a method invocation, a correspondent transition is created and attached to the CPN. If a loop control structure is found, it creates places to indicate the start and the end of the iteration and create transitions corresponding to methods invocations between them. When, an alternative structure control is found, the algorithm checks if there is trace line with IF and ELSE. Then, it creates two places labeled ALT | BEGIN | CONDITION and ALT | END that indicate the start and the end of the if else test. After that, it checks if there is a method invocation after the trace line IF | BEGIN, a transition with the same label is created then another transition with the method invocation. Otherwise, a transition with the label ELSE | BEGIN is created. In the case when only IF is found without ELSE the algorithm creates two places labeled OPT | BEGIN | CONDITION and OPT | END.

TABLE II. EXECUTION TRACE CORRESPONDING TO SCENARIO 1

```
Trace1 (nbr_article = 3, isInvoice = false, isPayslip = false):
L0. 1:BLL:Vendor:vendor  | Vendor () | DAL:VendorDAL:vendorDAL
L1. 1:BLL:Vendor:vendor  | newSales (nbr_article, isInvoice, isPayslip ) |
BLL:Sale:sale
L2. FOR |  BEGIN | i<=nbr_article
L3. 1:BLL:Sale:sale  | addArticle () | DAL:SaleDAL:saleDAL
L4. 1:BLL:Sale:sale  | addArticle () |  BLL:Sale: sale
L5. 1:BLL:Sale:sale  | calculAmount(newValue, oldValue) | BLL:Calcul:calcul
L3. 1:BLL:Sale:sale  | addArticle () | DAL:SaleDAL:saleDAL
L4. 1:BLL:Sale:sale  | addArticle () |  BLL:Sale:sale
L5. 1:BLL:Sale:sale  | calculAmount(newValue, oldValue) |  BLL:Calcul:calcul
L3. 1:BLL:Sale:sale  | addArticle () | DAL:SaleDAL:saleDAL
L4. 1:BLL:Sale:sale  | addArticle () | BLL:Sale:sale
L5. 1:BLL:Sale:sale  | calculAmount(newValue, oldValue)  | BLL:Calcul:calcul
L6. FOR | END
L7. IF |  BEGIN | isInvoice
L8. ELSE |  BEGIN
L9. 1:BLL:Sale:sale  | getDelivery () | BLL:Delivery:delivery
L10. 1:BLL:Delivery:delivery | signDelivery () |  BLL:Vendor:vendor
L11. 1:BLL:Delivery:delivery | getDelivery () | DAL: DeliveryDAL:deliveryDAL
L12. IF | END
L13   IF | BEGIN |  isPayslip
L14. IF | END
```

Scenario1

TABLE III. EXECUTION TRACE CORRESPONDING TO SCENARIO 2

```
Trace2 (nbr_article = 1, isInvoice = true, isPayslip = true):
L0. 1:BLL:Vendor:vendor | Vendor () | DAL:VendorDAL:vendorDAL
L1. 1:BLL:Vendor:vendor | newSales (nbr_article, isInvoice, isPayslip ) |
BLL:Sale:sale
L2. FOR | BEGIN | i<=nbr_article
L3. 1:BLL:Sale:sale | addArticle () | DAL:SaleDAL:saleDAL
L4. 1:BLL:Sale:sale  | addArticle () | BLL:Sale:sale
L5. 1:BLL:Sale:sale  | calculAmount(newValue, oldValue) | BLL:Calcul:calcul
L6. FOR | END
L7. IF | BEGIN | isInvoice
L15. 1:BLL:Sale:sale  | getInvoice() | BLL:Invoice:invoice
L16. 1:BLL:Sale:sale  | waitMessage ()  | BLL:Vendor:vendor
L17   PAR | BEGIN
L18. 10:BLL:Invoice:invoice  | preparingInvoice () | BLL:Invoice:invoice
L19. 10:BLL:Invoice:invoice  | signInvoice () | BLL:Vendor:vendor
L20. 10: BLL:Invoice:invoice | addArticle () | DAL:InvoiceDAL:invoiceDAL
L21. PAR | END
L8. ELSE | BEGIN
L12. IF | END
L13   IF | BEGIN |  isPayslip
L22. 1:BLL:Sale:sale  | getPayslip () |  BLL: Payslip: paySlip
L23. 1:BLL: Payslip: paySlip | signPayslip () |  BLL:Vendor:vendor
L24. 1:BLL: Payslip: paySlip | getPayslip () | DAL: PayslipDAL: paySlipDAL
L14. IF | END
```

Scenario 2

TABLE IV. EXECUTION TRACE CORRESPONDING TO SCENARIO 3

Trace3 (nbr_article = 4, isInvoice = false, isPayslip = true):	**Scenario 3**
L0. 1:BLL:Vendor:vendor | Vendor () | DAL:VendorDAL:vendorDAL	
L1. 1:BLL:Vendor:vendor | newSales (nbr_article, isInvoice, isPayslip) | BLL:Sale:sale	
L2. FOR | BEGIN | i<=nbr_article	
L3. 1:BLL:Sale:sale | addArticle () | DAL:SaleDAL:saleDAL	
L4. 1:BLL:Sale:sale | addArticle () | BLL:Sale:sale	
L5. 1:BLL:Sale:sale | calculAmount(newValue, oldValue) | BLL:Calcul:calcul	
L3. 1:BLL:Sale:sale | addArticle () | DAL:SaleDAL:saleDAL	
L4. 1:BLL:Sale:sale | addArticle () | v:Sale:sale	
L5. 1:BLL:Sale:sale | calculAmount(newValue, oldValue) | BLL:Calcul:calcul	
L3. 1:BLL:Sale:sale | addArticle () | DAL:SaleDAL:saleDAL	
L4. 1:BLL:Sale:sale | addArticle () | BLL :Sale:sale	
L5. 1:BLL:Sale:sale | calculAmount(newValue, oldValue) | BLL:Calcul:calcul	
L3. 1:BLL:Sale:sale | addArticle () | DAL:SaleDAL:saleDAL	
L4. 1:BLL:Sale:sale | addArticle () | BLL :Sale:sale	
L5. 1:BLL:Sale:sale | calculAmount(newValue, oldValue) | BLL:Calcul:calcul	
L6. FOR | END	
L7. IF | BEGIN | isInvoice	
L8. ELSE | START	
L9. 1:BLL:Sale:sale | getDelivery () | BLL:Delivery:delivery	
L10. 1:BLL:Delivery:delivery | signDelivery () | BLL:Vendor:vendor	
L11. 1:BLL:Delivery:delivery | getDelivery () | DAL: DeliveryDAL:deliveryDAL	
L12. IF | END	
L13 IF | BEGIN | *isPayslip*	
L22. 1:BLL:Sale:sale | getPayslip () | BLL: Payslip: paySlip	
L23. 1:BLL: Payslip: paySlip | signPayslip () | BLL:Vendor:vendor	
L24. 1:BLL: Payslip: paySlip | getPayslip () | DAL: PayslipDAL: paySlipDAL	
L14. IF | END	

The Places Pi and Pf are added to the CPN to indicate respectively the initial place and the final trace. To simplify the CPNs, all repeated method invocation between places “LOOP | BEGIN” and “LOOP | END” is deleted. All places representing trace lines are colored with the same color. These colors are used to diferentiat between the different scenarios.

After merging CPNs that refers to scenario1, scenario2 and scenario3, using the adapted Kbehavior, a new CPN is generated (Fig. 7). This CPN includes diferent paths with differents colors. Scenario1 has a yellow color, scenario2 has the green color while scenario3 has the red color.

The condition C1 refers to when the variable i is less than nbr_article while the condition C2 corresponds to if the variable isinvoice is true. Now, we apply our last transformation on loop places to make the obtained CPN more coherent (Fig. 8).

D. HLSD Extraction

The objective of this step is to extract the HLSD that represent the system behavior. For that, we use the transformation rules described in Section 4.4 to transform the final CPN into HLSD (Fig. 9).

The approach, as shown in following figure, is able to extract HLSD with the main UML2 fragment operators (seq,opt, alt and loop). Unlike other approaches using dynamic analysis, our approach succeeds in extracting the conditions corresponding to the combined fragment operators loop, opt and alt.

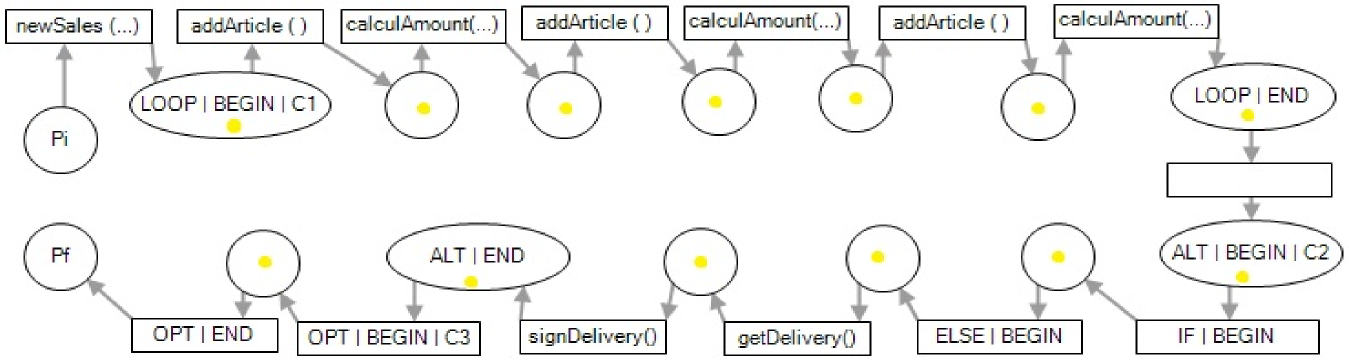


Fig. 4. CPN Corresponding to Scenario1.

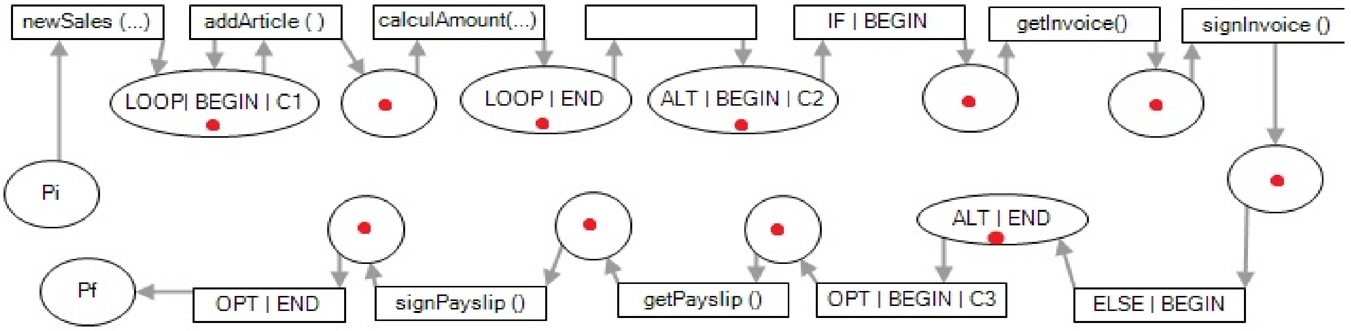


Fig. 5. CPN Corresponding to Scenario2.

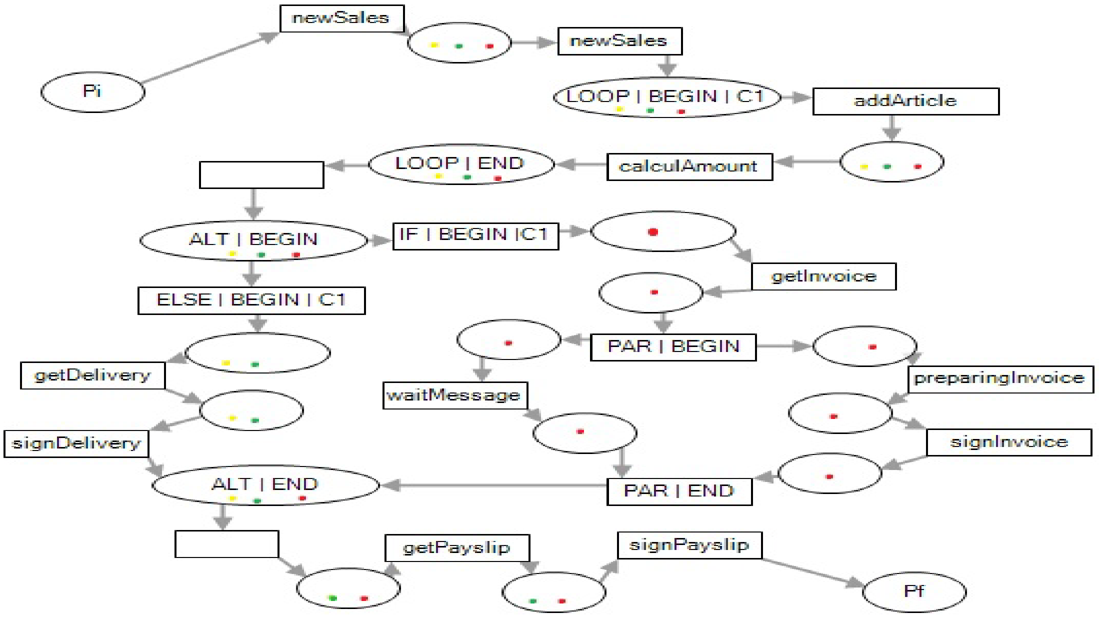


Fig. 6. CPN Corresponding to Scenario3.

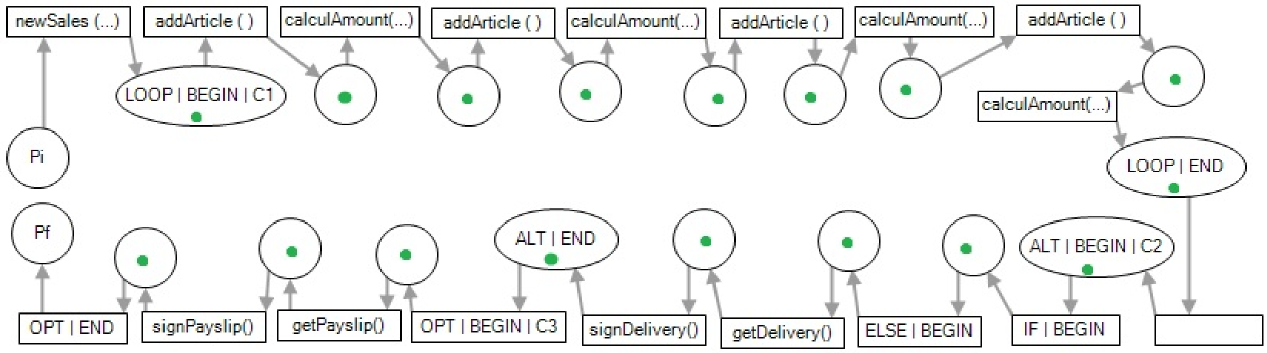


Fig. 7. The Merged CPN.

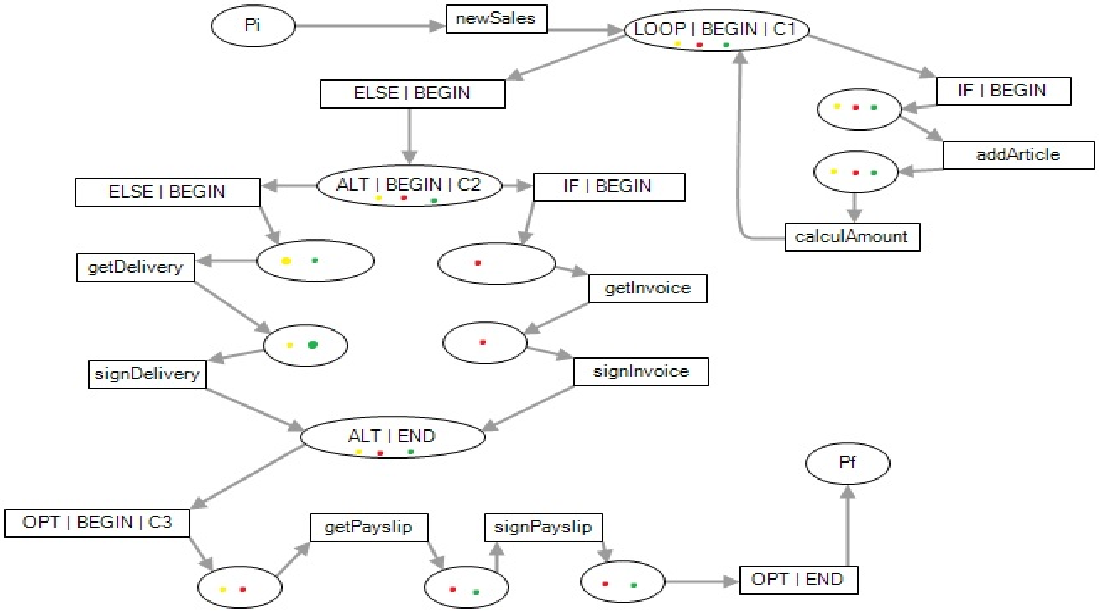


Fig. 8. The Finale CPN.

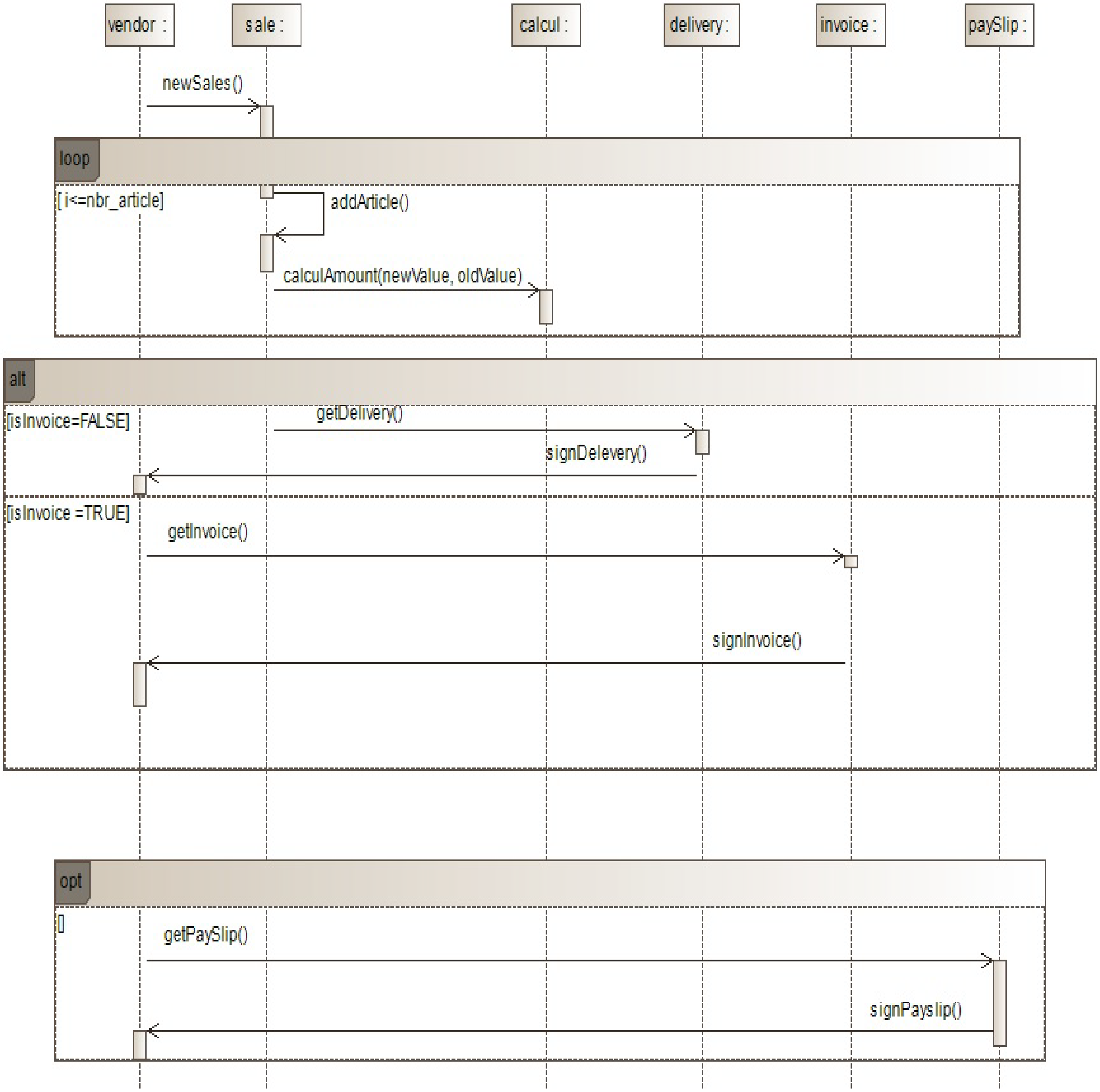


Fig. 9. Extracted HLSD.

In addition, our approach can be generalized to all object-oriented languages since it only uses text files for execution traces.

The colors are used to facilitate understinding the behavior of the system by subdividing it into several HLSD.

VI. Conclusion

Our work consists on proposing a new methodology for recovering an UML2 HLSD from execution trace using CPNs. For this, we first present a background of SD, CPN and reverse engineering. Then, we define several concepts which are essential for the understanding of our new approach. The approach starts by generating and collecting traces. Then, these traces are filtered and represented on CPNs in order to merge them. This merging is performed using an adapted version of the kBehavior algorithm that we have created. These CPNs are less complexes and more coherent than CPN in [8, 9]. The final obtained CPN use colors to differentiate between paths that represents different scenarios of the behavior of the system. This facilitates the understanding of the system. The approach succeeds to extract SD fragments operators such as seq, loop, alt and opt. It’s also extracts UML2 operator conditions relating on alt, opt and loop which is not the case in [8, 9].

Our future work is to extract the fragment operator *par* which is important for multi-threading systems. Besides, we will try to handle the problem of extracting others UML2 diagrams like a state diagram and activity diagram.

REFERENCES

[1] Sommerville, I., “Software Engineering” Addison Wesley, 2000.

[2] B. Cornelissen, A. Zaidman, et A. Deursen, “A Controlled Experiment for Program Comprehension Through Trace Visualization,” pp 2. IEEE Trans. on Software Engineering, 2011.

[3] K.-K. Lau and R. Arshad, “A Concise Classification of Reverse Engineering Approaches for Software Product Lines”,vol. 4, 2016.

[4] IEEE. std 1219: Standard for Software Maintenance. IEEE Computer Society Press, Los Alamitos, CA, USA, 1998.

[5] OMG. Unified Modeling Language (OMG UML), Superstructure, Vol. 2, 2007.

[6] L. C. Briand, Y. Labiche, J. Leduc, “Towards the Reverse Engineering of UML Sequence Diagrams for Distributed Java Software,” IEEE Transactions on Software Engineering, vol. 32, no. 9, pp. 642-663, 2006.

[7] C. Bennett, D. Myers, M.-A. Storey, D. M. German, D. Ouellet, M. Salois, and P. Charland, “A survey and evaluation of tool features for understanding reverse-engineered sequence diagrams,” J. Softw. Maint. E vol., vol. 20, no. 4, pp. 291–315, 2008.

[8] Chafik B., El Mahi B. and Abdeslam J.: A “Dynamic Analysis for Reverse Engineering of Sequence Diagram Using CPN,” Lecture Notes in Computer Science (ISSN: 0302-9743), 2018.

[9] Chafik B., El Mahi B., Abdeslam J. “A New Approach for Recovering High-Level Sequence Diagrams from Object-Oriented Applications,” Elsevier Procedia Computer Science Journal (ISSN: 1877-0509), 2019.

[10] E. J. Chikofsky and J. H. Cross, II, “Reverse Engineering and Design Recovery: A Taxonomy,” IEEE Software, vol. 7, no. 1, pp. 13-17, 1990.

[11] R. Kollmann and M. Gogolla, “Capturing Dynamic Program Behaviour with UML Collaboration Diagrams,” in Proceedings of the 5th Conference on Software Maintenance and Reengineering (CSMR’01), pp 58-67. IEEE Computer Society, 2001.

[12] R. Kollmann, P. Selonen, E. Stroulia, T. Syst¨a, and A. Z¨undorf. “A Study on the Current State of the Art in Tool-Supported UML-based Static Reverse Engineering,” in Proceedings of the 9th Working Conference on Reverse Engineering (WCRE’02), pp 22-32. IEEE Computer Society, 2002.

[13] A.Rountev, O. Volgin, and M. Reddoch, “Static Control-Flow Analysis for Reverse Engineering of UML Sequence Diagrams,” in ACM SIGSOFT Software Engineering Notes, ACM, vol.31, no.1, pp. 96-102, 2005.

[14] A. Rountev and B.H. Connell, “Object Naming Analysis for Reverse-Engineered Sequence Diagrams,” in Proceedings of the 27th International Conference on Software Engineering (ICSE’05), pp 254-263. ACM, 2005.

[15] Taniguchi, T. Ishio, T. Kamiya, S. Kusumoto, and K. Inoue, “Extracting Sequence Diagram from Execution Trace of Java Program,” International Workshop on Principles of Software Evolution (IWPSE’2005), pp. 148-151, 2005.

[16] Romain Delamare, Benoit Baudry, Yves Le Traon, “Reverse-engineering of UML 2.0 Sequence Diagrams from Execution Traces”, in Proceedings of the workshop on Object-Oriented Reengineering at ECOOP 06, 2006.

[17] Tewfik Ziadi, Marcos Aur’elio Almeida da Silva, Lom Messan Hillah, Mikal Ziane. “A Fully Dynamic Approach to the Reverse Engineering of UML Sequence Diagrams,” 16th IEEE International Conference on Engineering of Complex Computer Systems, ICECCS, Las Vegas, United States, 2011.

[18] B. Cornelissen, A. van Deursen, L. Moonen, and A. Zaidman, “Visualizing Test suites to Aid in Software Understanding,” In Proceedings of the 11th European Conference on Software Maintenance and Reengineering (CSMR’07), pages 213-222. IEEE Computer Society, 2007.

[19] K. Jensen, “A brief introduction to coloured Petri nets,” in Proceeding of the Tools and Algorithms for the Construction and Analysis of Systems (TACAS’97) Workshop, LNCS, Springer-Verlag, vol. 1217. pp. 203–208, 1997.

[20] A. Jakimi, A. Sabraoui, E. Badidi, A. Salah, and M. El Koutbi, “Using UML Scenarios in B2b Systems,” IIUM Engineering Journal, 2010

[21] AspectJ: The AspectJ project at Eclipse.org, http://www.eclipse.org/aspectj/.

[22] J. A. Brzozowski, “Derivatives of regular expressions,” J.ACM, vol. 11, no. 4, pp. 481–494, 1964.

[23] L. Mariani, F. Pastore and M. Pezze. “Dynamic Analysis for Diagnosing Integration Faults,” in IEEE Transactions on Software Engineering, vol. 37, no 4, pp. 486-508, 2011.

[24] A. Biermann and J. Feldmann,. “On the synthesis of finite state machines from samples of their behavior,” IEEE Transactions on Computer, vol. 21, pp. 592–597, 1972.

The Architecture of Intelligent Career Prediction System based on the Cognitive Technology for Producing Graduates to the Digital Manpower

Pongsaton Palee[1], Panita Wannapiroon[2]
Division of Information and Communication Technology
for Education, Faculty of Technical Education
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand

Prachyanun Nilsook[3]
Dept. Information and Communication Technology
for Education, Faculty of Technical Education
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand

***Abstract*—This research is a documentary research aimed at designing the architecture of the intelligent career prediction system based on the cognitive technology for producing graduates to the digital manpower. The research methods were divided into three phases: Phase 1, Composition Synthesis of Intelligent Career Prediction System. Phase 2, Intelligent Career Prediction System Architecture Designing based on the Cognitive Technology for producing graduates to the digital manpower. Phase 3, an assessment of the suitability of the architecture of the intelligent career prediction system based on the cognitive technology for producing graduates to the digital manpower. The architecture of the intelligent career prediction system by using the cognitive technology can be divided into three parts: 1) People involved in the architecture of the intelligent career prediction system consisting of five groups of related persons: students, staff, teachers, digital Enterprises, system administrator. 2) The architecture of the intelligent career prediction system consisting of four components: 1) User management, 2. Prediction Data Management, 3) Prediction Management system, 4) Prediction Display system, and cloud computing, an assessment of the suitability the architecture of the intelligent career prediction system based on the cognitive technology for producing graduates to the digital manpower by nine experts in the intelligent career prediction system and cognitive technology. The statistics used in the research are Mean and standard deviation. The evaluation results showed that the developed architecture was the most suitable, with the combined mean of 4.54, and the standard deviation was 0.49.**

Keywords*—Architecture of intelligent career prediction system; cognitive technology; producing graduates; digital manpower

I INTRODUCTION

Education is the process of developing people in society to be competent, the educational system is an important factor for the growth and development of the country, as information technology and digital technology has changed and developed more and more, the world is moving towards In the age of digital manpower generation, educational institutions, especially universities, play an important role in producing graduates that meet the modern labor market needs, universities need to adapt still include: 1) Curriculum development is up to date in modern times and more important to keep up with the technology that changes the way of life of the people in society. 2) Technology, teaching and learning tools, practical teaching for students to actually do their work, learn to work as a team and to solve immediate problems. 3) Educational personnel must act as coaches, provide guidance in order to make students the center of learning. 4) Universities should not only have a network of educational institutions but must build a network of partnerships with private businesses in order to create opportunities for students [1],[2].

Building the architecture of the intelligent career prediction system based on the cognitive technology for producing graduates to the digital manpower, to create a system for predicting further education of students who will enter the university, and also guide students who are about to graduate. Educators who are interested in pursuing a career in the digital Manpower [3], provide planning and preparation for such careers, resulting in the university developing a modern curriculum to accommodate the digital Manpower of Standardized, Reliable, Flexible Management of Learning Objectives, especially the intelligent career prediction system that can analyze digital Manpower professions, and also analyze what competencies students need in such professions to support the digital manpower in the upcoming Eastern Economic Corridor [4],[5] .

This architecture of the intelligent career prediction system based on the cognitive technology for producing graduates to the digital manpower presents a conceptual framework for the development of an information system model that combines the characteristics of the intelligent career prediction system with the cognitive technology to develop the next genius career prediction system.

II RESEARCH OBJECTIVES

1) To synthesize the compositions of the architecture of the intelligent career prediction system based on the cognitive technology for producing graduates.

2) To design the architecture of the intelligent career prediction system based on the cognitive technology for producing graduates to the digital manpower.

3) To assess the suitability of the architecture of intelligent career prediction system based on the cognitive technology for producing graduates to the digital manpower.

III Related Works

Reference [6] The Digital career prediction system is the use of statistical algorithms data, and machine learning techniques to identify the likelihood of future results from historical data, predictive models use known results to develop or train. A model that can be used to predict values for different or new data, (for example, in our system, the target variable is a student's success in courses based on the estimated importance from a set of input variables, which is different from the descriptive model that helps in understanding what happened or a diagnostic model that helps in understanding important relationships and determining why and what happened.

Author(s) in [7],[8] Cognitive technology is the use of computers to simulate human learning, using cognitive technology learning algorithms to predict future results or to make different types of decisions under certain conditions. This allows the machine to understand the current situation on the basis of learning, enabling the machine to make informed decisions. Predictions, outcomes, or decisions will evolve as new information is received. Learning algorithms can be divided according to the learning process, the algorithm in the grouping (Classification) for learning to solve problems effectively (Supervised Learning) [9],[10]. The algorithm in grouping the exam level is divided into two steps: learning to create a template and the leveling of the exam to be graded, tested against an exam like a Training Data set, a process that uses the most common or similar search method to predict a data set that can be divided in any type of information. The Performance Model Assessment is the assessment of the model's ability to grade an exam's cognitive rating, through machine learning focused on decision-making or prediction, correct grading, a test method for comparing the model's performance to the performance of the exam's Cognitive rating, determined by its accuracy, from the Cognitive Rating [11].

The author in [12] Cloud computing is defined as a computing service or computer resource, covering the use of processors, memory, storage, and online systems from users to simplify the installation, administration, save time, reduce the cost of building the computer system, and the network itself. It is a behavior by using resources through the Internet. It is a processing method based on user needs. The user can specify a requirement to the system software, then the software requests the system to allocate resources and services to meet the user needs, which the system can increase and decrease the number of resources. Including offering services that fit the needs of users all the time, users can access various information systems via the Internet, users can manage system resources through the network, if the demand is greater, they can purchase additional services to increase their potential. The system without upgrading systems and computers, reduces costs and simplifies system administration and maintenance. The adoption of the intelligent career prediction system based on the cognitive technology on cloud computing to help to reduce costs, reduce time, simplify IT management, and when the demand grows, such as more users, better processing, more storage space, which can easily and conveniently expand the various resources of the Cloud system.

IV Proposed Methodology

The research method was divided into three phases according to the research objective.

Phase 1. Intelligent career prediction system model synthesis is the study of information about the composition synthesis of the intelligent career prediction system, the researcher has studied the document, the content analysis, and has synthesized nine related researches.

Phase 2: The Designing of Intelligent Career Prediction System Architecture Designing based on the Cognitive Technology for Producing graduates to the digital manpower, Intelligent Career Prediction System Architecture based on the Cognitive Technology for Producing graduates to the digital manpower by logical design to provide a model of the architecture of Intelligent career prediction system based on the Cognitive Technology for Producing graduates to the digital manpower, which incorporates the concept of architecture development (Conceptual Framework).

Phase 3: An assessment of the suitability of the architecture of the intelligent career prediction system based on the cognitive technology for producing graduates to the digital manpower is an assessment of suitability by using nine qualified persons.

V Experimental Results and Discussions

Phase 1: Results of composition synthesis of the intelligent career prediction system by using the research to synthesize the components of the intelligent career prediction system using the synthetic results table as shown in Table I.

Phase 2: The Architecture of the Intelligent career prediction system by using cognitive technology.

The architecture of the intelligent career prediction system by using the cognitive technology could be divided into three parts: 1) People involved in the intelligent career prediction system, 2) The architecture of the intelligent career prediction system, 3) Cloud computing as shown in Fig. 1 and Fig. 2.

A. People involved in the Architecture of the Intelligent Career Prediction System.

People involved in the architecture of the intelligent career prediction system consisted of five groups of stakeholders: students, staff, digital Enterprises, universities, administrators, with each group responsible as shown in Table II.

TABLE I. SYNTHESIS TABLE OF THE INTELLIGENT CAREER PREDICTION SYSTEM MODEL FROM TABLE I, IT FOUND THAT THE INTELLIGENT CAREER PREDICTION SYSTEM CONSISTED OF 4 MAIN COMPONENTS

Composition of the Architecture of Digital Career Prediction System	[13] Bostandjiev et al, 2013	[14] Heap et al, 2014	[15] Wang et al, 2014	[16] Razak et al, 2014	[17] Gupta and Garg, 2014	[18] Yeh and Chen, 2015	[19] Kaensar , 2015	[20] Liu et al., 2016a	[21] Liu et al., 2016b	Synthesis Results
1. Manage Users	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2. Data Management	✓	✓	✓	✓	✓	✓		✓	✓	✓
3. Analysis Management	✓		✓		✓	✓		✓		
4. Management System	✓	✓	✓	✓	✓	✓		✓	✓	✓
5. Storage system		✓	✓				✓			
6. Evaluation Management	✓		✓	✓			✓			
7. Suggestion	✓	✓		✓		✓		✓	✓	
8. Display system	✓	✓	✓	✓	✓	✓		✓	✓	✓

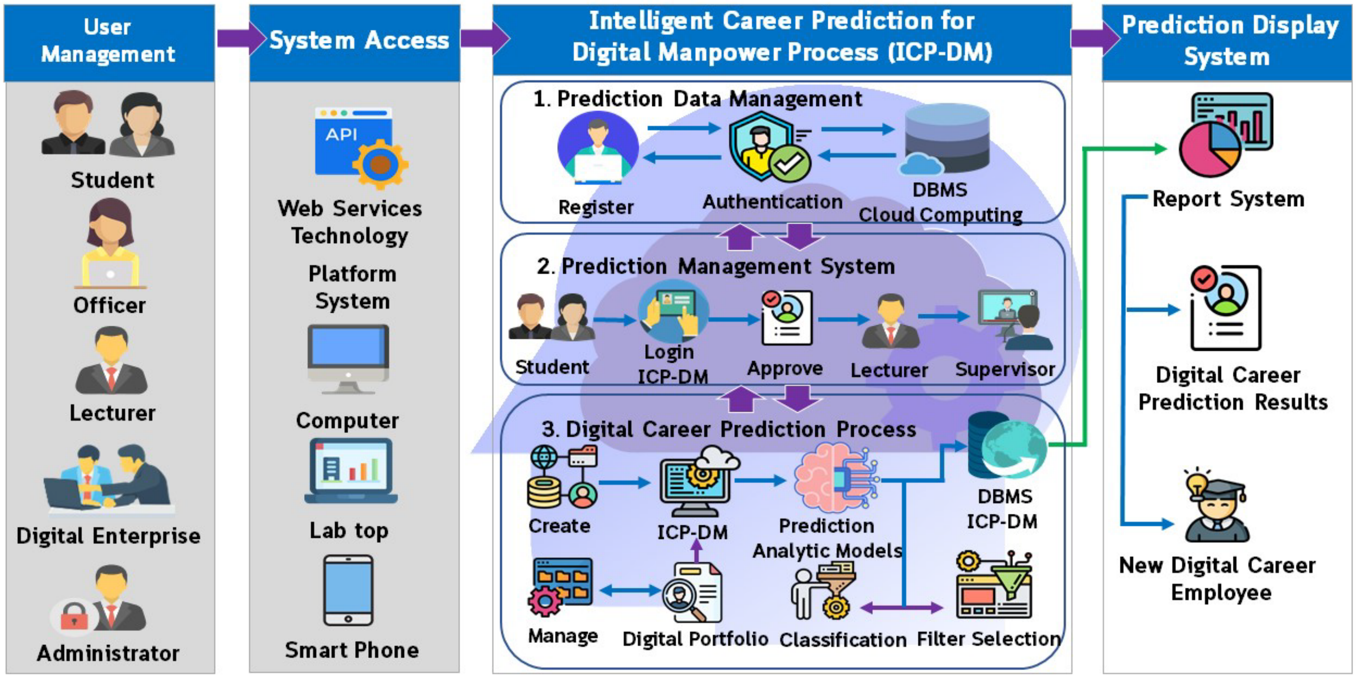


Fig 1. The Architecture of the Intelligent Career Prediction System based on the Cognitive Technology for Producing Graduates to the Digital Manpower.

TABLE II. PERSONS INVOLVED IN THE ARCHITECTURE OF THE INTELLIGENT CAREER PREDICTION SYSTEM

Persons involved in the architecture of the intelligent career prediction system	**Responsibility**
1. Student	- Personal Information Management, grades, experience data, education, digital portfolios, other data generated by the intelligent career prediction system. - Check the results of the selection to work - Check the Enterprise's advice.
2. officer	- Prepare career talent information - Prepare job position information - Check the information for accuracy
3. Teacher	- Prepare career talent information - Prepare job position information and check the information for accuracy
4. Digital Enterprise	- Check the results of the selection to work - Check the information of job applicants.
5. Administrator	- User Account Management - License Permission - Other job management

B. The Architecture of the Intelligent Career Prediction System the Architecture of the Intelligent Career Prediction System Consisted of the main Modules and Sub Modules as Shown in Table III

Phase 3: An Assessment of the suitability the architecture of the intelligent career prediction system based on the cognitive technology.

The Analysis of the suitability of the architecture of the intelligent career prediction system, the cognitive technology for producing graduates to the digital manpower by using descriptive statistics such as Mean and Standard Division, setting criteria for assessing suitability as Rating Scale Model, which has the criteria for determining the weight of the assessment into five levels according to the Linkert's Scale with the following criteria in Table IV.

- 5 represents the most suitable
- 4 represents very suitable
- 3 represents moderate suitable
- 2 represents less suitable
- 1 represents the least suitable

The Interpretation criteria to categorize the average score of suitability of the experts, there are scoring criteria for each level as follows:

TABLE III. THE MODULE OF THE ARCHITECTURE OF THE INTELLIGENT CAREER PREDICTION SYSTEM

Main Modules	**Sub Modules**	**Responsibility**
1. User Management	1.1 Register	Username and Password First time registration, user login, user must enter Username and Password.
	1.2 Edit Profiles	Members change information, personal details such as changing password, email address, profile picture.
	1.3 License Permission	License Permission to different types of users such as students, or administrators.
2. Prediction Data Management	2.1 Selection Prediction Methods	Show identification values for career prediction.
	2.2 Cognitive Technology	Use the cognitive technology to select the rules of divination according to the conditions.
3. Prediction Management system	3.1 Evaluation Examination	Show a selection of several career prediction methods.
	3.2 Selection input file processing	Show a selection of files, import career data to predict.
	3.3 Cognitive Technology	Use the cognitive technology for choosing the right predictions for the profession.
4. Prediction Display System	4. View Position Digital Career	Show the results of job titles or occupation groups recommended to users.
	4.2 Report Digital Career statistics	Show student statistics, employment status of graduates in 3D graph format.

TABLE IV. ARCHITECTURAL EVALUATION CRITERIA USING LINKERT'S SCALE

Mean frequency value	Meaning of values
4.50-5.00	represents to the most suitable
3.50-4.49	represents that is very suitable
2.50-3.49	represents moderate suitable
1.50-2.49	represents less suitable
1.00-1.49	represents the least suitable

An assessment of the suitability of experts in the intelligent career prediction system and the cognitive technology towards the architecture of the intelligent career prediction system by using the cognitive technology, the researcher had nine experts to do the assessment, and the suitability assessment results could be shown in Table V.

Table V shows Assessment Results of the suitability of the architecture of the intelligent career prediction system by using the cognitive technology.

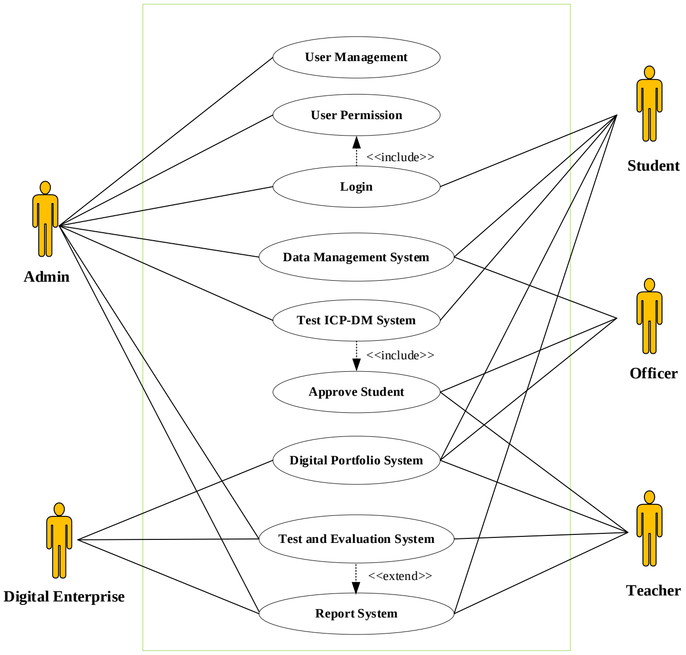


Fig 2. The Use Case Diagram of the Intelligent Career Prediction System User based on the Cognitive Technology for Producing Graduates to the Digital Manpower.

TABLE V. ASSESSMENT RESULTS OF THE SUITABILITY OF THE ARCHITECTURE OF THE INTELLIGENT CAREER PREDICTION SYSTEM BY USING THE COGNITIVE TECHNOLOGY

Details of the model	**Assessment results**		**Suitability level**	**Level**
	$\overline{X}$	**S.D.**		
1. Principles and concepts used as the basis for the development of the architecture of the intelligent career prediction system based on the cognitive technology for graduate production.	4.55	0.48	Most suitable	
2. The architecture of the intelligent career prediction system based on the cognitive technology				
2.1 People involved in the intelligent career prediction system.				
2.1.1 Students	4.60	0.55	Most suitable	3
2.1.2 Teacher	4.62	0.51	Most suitable	2
2.1.3 Officer	4.77	0.44	Most suitable	1
2.1.4 Digital Enterprise	4.44	0.33	Highly suitable	5
2.1.5 Administrators	4.55	0.48	Most suitable	4
2.2 The components of the intelligent career prediction system based on the cognitive technology.				
2.2.1 User Management	4.55	0.48	Most suitable	3
2.2.2 Prediction Data Management	4.38	0.51	Highly suitable	4
2.2.3 Prediction Management system	4.33	0.55	Highly suitable	5
2.2.4 Prediction Display system	4.80	0.45	Most suitable	1
3. Cloud computing	4.60	0.55	Most suitable	2
4. The architecture of the Intelligent career prediction system based on the cognitive technology for producing graduates to the digital manpower that can be used for practical purposes.	4.33	0.55	Highly suitable	
Summary of assessment items	**4.54**	**0.49**	Most suitable	

From Table V, assessment Results of the suitability of the architecture of the intelligent career prediction system were overall at the most level, with the mean of 4.51 and the standard deviation was 0.49, which was in the range 0 to 1, could be considered reliable data so it could be concluded that the experts with relatively similar opinions were most suitable.

VI Discussion

This research is a documentary research, divided into two phases: component synthesis of the intelligent career prediction system and architectural design of the intelligent career prediction system by using the cognitive technology, data analysis by using content analysis techniques, research findings were in Fig. 1 and Fig. 2.

- The Intelligent career prediction system consisted of four components: 1) User management, 2) Prediction Data Management, 3) Prediction management system, 4) Prediction display system.
- The architecture of the intelligent career prediction system by using the cognitive technology could be divided into two parts: 1) those involved in the intelligent career prediction system, and 2) the architecture of the intelligent career prediction system consisted of four main modules: 1) User Management module consisted of 3 sub-modules: 1.1 Login module. (Register), 1.2 Edit Profiles Module, 1.3 License Permission. 2) Prediction Data Management module consisted of two sub-modules: 2.1 Selection Prediction methods, 2.2 Using the cognitive technology to select prediction rules to meet the conditions. 3) Prediction Management system module consisted of 3 sub-modules: 3.1 Evaluation data. Examination, 3.2 Selection input file processing, 3.3 Using the cognitive technology for selecting suitable prediction results for the profession. 4) Prediction display system module consisted of 2 sub-modules: 4.1 View position Digital Career, 4.2 Report Digital Career statistics, and then analyzed and designed the system and information for further research work.

VII Result

Assessing the suitability of the architecture of the intelligent career prediction system based on the cognitive technology for producing graduates to the digital manpower by nine experts in the intelligent career prediction system and the cognitive technology, the statistics used in the research are Mean and Standard Deviation, the Assessing results showed that the developed architecture was the most suitable, with the total mean of 4.51 and the standard deviation of 0.49. The Cognitive Technology Classification processes uses data mining techniques to automatically group the items according to Bloom's taxonomy, which is consistent with the research of [22], [23], [24], [25] and [26] that grouped and classified based on Bloom's Revised Taxonomy.

Acknowledgment

Thanks to the experts in computer technology and digital manpower who have provided assistance in assessing, certifying, and providing suggestions for improving develop the Structural Equation Modeling for Digital Career Prediction Factor , thanks to National Research Council of Thailand, for supporting funding for this research, and Thanks to the advisor from King Mongkut's University of Technology North Bangkok, That provides advice and guides in organizing research in this time to be successful.

References

[1] E-government office (2017). (Draft) The Digital Government Development Plan of Thailand 2017 - 2021. First Edition Bangkok: Bolliger & Company (Thailand).

[2] Institute of professional qualifications (Public Organization) 2017. Occupational Standards Database, Professional Qualifications Software and Applications field. Retrieved from: https://tpqi-test.tpqi.go.th/tpqi_sa/index.phppage=Pathway.php&OCC=SWA.

[3] World Economic Forum (2016). "Global Challenge Insight Report -The Future of Jobs Employment, Skills and Workforce Strategy for the Fourth http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf.

[4] Eastern Economic Corridor Development Project. (2017) Action Plan for Education, Research and Technology Personnel Development. Retrieved from https://www.eeco.or.th.

[5] Ministry of Digital Economy and Society. (2016). Digital Economy and Society Development Plan. First edition, Bangkok:(Electronic document). Retrieved 14 October 2020. From https://www.dga.or.th/th/profile/2008.

[6] Kalyankar, G. D., Poojara, S. R., & Dharwadkar, N. V. (2017). Predictive Analysis of Diabetic Patient Data Using Cognitive Technology and Hadoop.2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2017), 619–624. https://doi.org/10.1109/I-SMAC.2017.8058253.

[7] F. Mart, L. Contreras-ochando, M. Kull, and N. Lachiche, "CRISP-DM Twenty Years Later : From Data Mining Processes to Data Science Trajectories," vol. 4347, no. c, 2019, doi: 10.1109/TKDE.2019.2962680.

[8] W. Puarungroj, N. Boomsirisumpun, P. Pongpatrakant, and S. Phromkot, "A preliminary Implementation of Data Mining Approaches for Predicting the Results of English Exit Exam," in *International Conference on Information Technology (INCIT)*, 2017, vol. 53, no. 9, pp. 1689–1699, doi: 10.1017/CBO9781107415324.004.

[9] C. Wongwatkit, "An Online Web-based Adaptive Tutoring System for University Exit Exam on IT Literacy," *Int. Conf. Adv. Commun. Technol. ICACT*, vol. 2019-Febru, pp. 563–568, 2019, doi: 10.23919/ICACT.2019.8701994.

[10] J. Flowers, "The problem in technology education (A definite article)," *J. Technol. Educ.*, vol. 21, no. 2, pp. 10–20, 2010, doi: 10.21061/jte.v21i2.a.2.

[11] N. Ketui, W. Wisomka, and K. Homjun, "Using classification data mining techniques for students performance prediction," ECTI DAMT-NCON 2019 - 4th Int. Conf. Digit. Arts, Media Technol. 2nd ECTI North. Sect. Conf. Electr. Electron. Comput. Telecommun. Eng., pp. 359–363, 2019.

[12] Akrima Boonyoo and Nopadon Kaewbanphot. 2015. What is *Cloud computing*? Library and Information Center, Science and Technology Department of Science Service. Retrieved from: http://lib3.dss.go.th/fulltext/dss_knowledge/bsti-6-2558-cloud.pdf.

[13] Bostandjiev, S., O'Donovan, J., Höllerer, T., 2013. LinkedVis: exploring social and Semantic carrer recommendations. Proceedings of the 2013 International. Conference on Intelligent User Interfaces. ACM, pp. 107–116.

[14] Heap, B., Krzywicki, A., Wobcke, W., Bain, M., Compton, P., 2014. Combining career progression and profile matching in a job recommender system. In: Pacific Rim International Conference on Artificial Intelligence. Springer International Publishing, pp. 396–408.

[15] Wang, Y., Zhang, X., Nan, L., Wang, D., 2014. Digital Career recommendation based on student achievement mining in vocational skill training. In: 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). IEEE, pp. 686-691.

[16] Razak, T.R., Hashim, M.A., Noor, N.M., Halim, I.H.A., Shamsul, N.F.F., 2014. Career path recommendation system for UiTM Perlis

students using fuzzy logic. In: 2014 5th International Conference on Intelligent and Advanced Systems (ICIAS), IEEE.

[17] Gupta, A., Garg, D., 2014. Applying data mining techniques in job recommender system for considering candidate job preferences. In: 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, pp. 1458–1465.

[18] Yeh, C.T.D., Chen, T.M., 2015. Ubiquitous job recommendation system for graduates in Taiwan. Int. J. Electron. Commerce Stud. 6 (1), 127–136.

[19] Kaensar, C., 2015. Design of personal job recommendation framework on smartphone platform. World Acad. Sci. Eng. Technol. Int. J. Comput. Electr. Autom. Control Inf. Eng. 9 (12), 2315–2319.

[20] Liu, R., Ouyang, Y., Rong, W., Song, X., Xie, W., Xiong, Z., 2016a. Employer oriented recruitment recommender service for university students. In: International Conference on Intelligent Computing, Springer International Publishing.

[21] Liu, R., Ouyang, Y., Rong, W., Song, X., Tang, C., Xiong, Z., 2016b. Rating prediction based job recommendation service for college students. In: International Conference on Computational Science and Its Applications. Springer International Publishing, pp. 453–467.

[22] C. Paiwithayasiritham, P. Makmee, and K. Mingsiritham, "The Development of a training program for basic education school teachers on developing the higher level learning assessment methods," *Veridian E-Journal, Silpakorn Univ*., vol. 6, no. 1, pp. 904–913, 2013.

[23] U. Shafique and H. Qaiser, "A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)," *Int. J. Innov. Sci. Res*., vol. 12, no. 1, pp. 217–222, 2014.

[24] C. Paiwithayasiritham, P. Makmee, and K. Mingsiritham, "The Development of a training program for basic education school teachers on developing the higher level learning assessment methods," *Veridian E-Journal, Silpakorn Univ*., vol. 6, no. 1, pp. 904–913, 2013.

[25] M. H. Ying, S. H. Huang, and L. R. Wu, "An item selection strategy based on association rules and genetic algorithms," *2009 4th Int. Conf. Innov. Comput. Inf. Control. ICICIC 2009*, pp. 1040–1044, 2009.

[26] M. Mohammed and N. Omar, "Question classification based on Bloom's Taxonomy using enhanced TF-IDF," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 8, no. 4–2, pp. 1679–1685, 2018.

Analyzing the Barriers and Possibilities with *p*-values towards Starting a New Postgraduate Computer and Engineering Programs at Najran University: A Cross-Sectional Study

Abdullah Alghamdi
Department of Information Systems
Najran University
Najran, The Kingdom of Saudi Arabia

***Abstract*—A cross-sectional study was conducted to find out the barriers and their possible solutions to start a new postgraduate computer and engineering programs at Najran University (NU), Kingdom of Saudi Arabia. This study includes interviews and surveys consist of 35 questions. The total number of the participant was 363; most of them were employees at the government and private sectors. In this study to analysis the result IBM's Statistical Package for the Social Sciences (SPSS) version 22 is used to analyze the result with the respective *p*-values calculated using the Pearson Chi-square test. Study reveals that 95.6% of participants want to pursue a graduate degree. However, only (46.9%) can communicate in English academically. Among the respondents, about (42.1%) started a graduate program before, but only (11%) has completed the program. Others could not continue their graduate programs because they were not able to attend the classes coming from a far distance and uncomfortable classes schedule and time. Hence, questions were distributed among the participants to get their opinions to find the solution. Study reveals that among the participants' both government and private sectors employees shown the importance to have online classes with (69.56%) and to have a comfortable schedule and time of courses with (95.65%). This study outlined the main barriers and possible solutions and recommendations that may be helpful for higher education institutions and organizations to start new graduate programs.**

Keywords*—Postgraduate program; p-values; barriers; Najran university; English language; schedule

I Introduction

Starting a new graduate program in computer science and engineering is a complicated process and it requires a lot of strategic planning, curriculum building, and dedication of resources. This process gets even more complicated when starting the new program in a public, and government funded university. While going through this process, the author concluded that it is very important to document and share this experience.

As a newly appointed head of information systems department, and with average and minimum expertise in academia and management respectively, the author realized the need to have a postgraduate program in the department.

This program would be the first in computer science and engineering colleges of Najran University (NU), a 12 years old and far southern rural government funded university in Saudi Arabia. The computer science and engineering colleges are the twin colleges of the university starting with one dean for both colleges, a mathematician who grounded the same rules and set the directions for both colleges. In addition, both colleges have the same number of departments and about the same number of students and graduates every semester. Opening a new postgraduate program would be the first step toward a new vision for the computing and engineering majors in NU which in result would shift the focus to research rather than being on education only.

The author in this paper documented the experience in opening a new postgraduate major. The organization of this paper as follow: Section II: the author offers a literature review on computer education in general, and Section III: the author states the methodology followed in this research. After that, the results are presented in Section IV. Then, challenges and recommendations are presented in Section V, while the conclusion is stated in Section VI.

II Related Work

Developing and managing graduate programs is a complex process. In 2000, Eriksson has the chance to develop the guidelines for the doctoral programs in Malardens Hogskola University, a small university in Sweden [3]. Their focus was on how to develop a graduate education that maintains quality standards of Swedish education, while taking the findings of subcommittee of the association of universities in Sweden (SUHF) postgraduate study [3] into account. They invited all departments to propose postgraduate programs. In their first round, two proposal were rejected due to the clarity of the proposal, shape of the department made that proposal, and insufficient resources. The study also analyze the process of designing: syllabus, joint courses, individual study plans, supervisor selections, thesis defense mechanism. On the other hand, Cullen et al. studies the ways of promoting effectiveness in Ph.D. supervision in Australian universities [4]. The authors recommend that the students should have access to supervision panel rather than one supervisor which in result would provide the students with broader range of skills and expertise. Among other recommendations, the authors emphasized that students should be concentrated in groups of sufficient numbers, professor professional development should be provisioned, and a period of structured induction should be introduced.

The facilitators and barriers to attaining a postgraduate degree in South Africa are discussed in [5] where, lack of financial support, family commitments and lack of time are found the mentionable barriers among the postgraduate students. Whereas, students who had obtained a postgraduate degree thought one can facilitate through the degree by gaining of expertise, the fulfilment of a personal goal (personal development) and improvement of patient care (career perspective) as participants. Students with lack of motivation found one of the major barriers for not obtaining a postgraduate degree. The rate was significantly more likely (p < 0.05) to report [5]. Not enough funding [6], inflexibility of available programs [7], family commitments and geographical location [8], lack of time, personal motivation and recognition in terms of increased pay or promotion [9] are major causes and additional barriers to not pursuing or not completing postgraduate studies. Authors in [10] mentioned that a lack of training on academic writing and low publication rate in undergraduate programs poses a barrier to not starting postgraduate physiotherapy program. Almost a similar problem found in Kuwait-based physiotherapists reporting publications in the academic journal [11].

Some research uncovers several themes in the students' motivations for doing a postgraduate degree include career development motives, academic motives, social environment self-formation motives, knowledge and awareness, academic considerations, financial cost and the influence of peer norms [12]. Other obstacles implicating in not conducting higher studies and research include inadequate resources and facility for research, lack of concentration by faculty or guide, and unavailability of required data or samples [13]. Other key challenges are cultural barriers and language proficiency issues mentioned in [14] as the major obstacles to pursuing graduate studies. Researchers in [15] suggest that blended learning, a relatively new concept in Saudi Arabia, shows promising results with higher student satisfaction. Blended learning is an active and passive student-centered learning that may enhance critical thinking and application. Learning experiences can be enriched by adopting a blended method of instruction at various stages of undergraduate and postgraduate education. Postgraduate students need to conduct more research to prepare them to venture into business as an entrepreneur [16]. Authors in [17] identified specific features of postgraduate medical education innovation that include designing the curriculum with the needs of the users in mind, the challenge of implementing other competencies than that of the medical expert; organize educational support. They also emphasize on the importance of regional and inter-organizational implementation strategies, curriculum, training and educational support; facilitate knowledge sharing; buildup inter-organizational networks and cooperation [17].

A thorough investigation is done by Symons to study the transition to higher degrees and the difficulties faced by the students who are beginning coursework postgraduate programs [18]. The difficulties include the worry about expectations, time spent away from school in the workforce and encountered issues when coming back to school. Also, the higher standards and studying a discipline that is different from their undergraduate. The author recommends a good induction and training programs to help students have a satisfactory start, which is similar to what Boyle and Boice [19], and Youngman [20] have concluded. Interestingly, MacKay[21], Golde[22], Jurgs et al. [23] are all agrees that there is a fundamental assumption among many schools that only the imperfect and unpleasant students are experiencing problems. This assumption leads to the conclusion that the student selection process in accepting students into a graduate school is the solution for all the issues that face the students who are beginning a graduate program. However, those studies [21-23] shows that assumption is inaccurate.

Padro et al. provide a comprehensive overview of academic leadership, doctoral education, and student-supervisor relationship [24]. Also, the authors examine the external pressures including neoliberalism, accepting more students, quality benchmarking that faces the universities nowadays with focus on how to administer and deliver postgraduate programs. Besides, the authors provide recommendations to meet the organizational and students in many areas including professional, employment, information literacy, and processes of examinations.

The above information presents a paradox of sorts. Postgraduate studies appear to be the most efficient means of achieving growth in the profession-specific evidence. However, based on several reported barriers, globally there is a reluctance to pursue postgraduate studies. This study aims to investigate the potential limitation and provide recommendations to pursue postgraduate studies in Saudi universities, particularly in Najran University. It is hoped that the results of this study may be of benefit to policymakers and stakeholders in Najran University and other higher education providers to provide postgraduate studies for the students locally, regionally and globally.

III METHODOLOGY

The methodology followed in this paper includes interviews, a survey distributed to a sample of interested parties, and investigations of rules and regulations of the university and education laws in the Saudi government. The questionnaire includes about 35 questions that cover the general information of the respondents, their education, and their relative education status. Also, it includes questions about the bachelor's degree majors, grades, willingness to start a postgraduate program, and preferred majors. According to the latest official population count, Najran region has about 570,000 people, while Najran city has 340,000 people [1]. The target of this study is the people who majored or has the possibility to major in computing and engineering fields. There is no exact number of the population size of this study, however, with educated predication, the author estimates the population size to be between 30000 – 50000 people. The prediction is based on the number of students entering computer science and engineering majors in last ten years, the number of students in high schools in Najran region, and estimated number IT and engineering professionals in government and private sectors organizations. The valid responses received from respondents are 363 responses. The confidence level is 95% with a margin of error of 5%. In this study, the interviews conducted with current undergraduate students, faculty members, and officials from the city.

IV Major Findings and Results Analysis

Table I presents the demographic characteristics of the participants. Most of the respondents in this study are males (91.7%), while the females are about (8.3%). The reason for this is that among all the departments in college of computer science and college of engineering, there is only two departments are available for female students, one of them started at Fall 2019. Many women lack the interest of participation in the study because they have not majored in majors of interest in this study. The majority of the respondents are between the age of 18 and 40 years old with about (84.3%). When asked about the latest degree the participant achieved, only (11%) has achieved a graduate degree including higher diploma, masters, or PhDs.

Table II evaluates participants' interest in pursuing a graduate program at Najran University (NU). When the participants are asked about the city they live in currently, about (95%) live in Najran city and Najran region, which has an area of 365 Km^2 [2]. Aseer and Jazan regions are respectively 300Km, 650km away from the capital of Najran region, Najran city. This question is included in this section to emphasize that there is no graduate program in IT and engineering in this whole region, and the nearest programs would be in King Khaled University (KKU), which is about 300km away. Some of the participants in this study are majored in art, science, administrate and religion majors; however, the majority (60%) are from IT and engineering majors. (95%) of the participants are expecting to pursue their graduate degrees. Because the native language in Najran is Arabic and instruction in IT and engineering fields is in English, the participants were asked about their ability to communicate in English. Only (46.9%) are able to academically communicate in English. An interesting result is that about (70%) of the participants has started a graduate program sometime in the past. In Table I, the participants who finished their graduate degree are only (11%). One reason for that is that there are many people who start their program; however, they drop out before they finish. One interviewee said he dropped out because it's was hard for him to travel every week to Aseer region to attend classes, while some others dropped because they could not study and work at the same time. Very significant number are expecting to pursue their graduate degrees, which is an interesting news for NU.

TABLE I. Demographic Characteristics of Participants In The Study

Variable	**n (%)**
Gender	
Male	333 (91.7)
Female	30 (8)
Age	
less than 18	3(0.8)
18-24	80(22)
25-30	86(23.7)
31-40	140(38.6)
Above 40	54(14.9)
Nationality	
Saudi Citizen	354(97.5)
Non-Saudi	9(2.5)
Occupation	
Government	178(49)
Private sector	75(20)
Students	69(19)
Unemployed	36(10)
Retired	5(2)
Latest Degree	
High school	66(18.1)
Associate degree	51(14)
Bachelor	206(56.8)
Graduate Degree	40(11)
Has a relative with graduate degree	
Yes	238(65.5)
No	125(34.4)

TABLE II. Evaluation of Participants' Interest to Pursuing A Graduate Program at NU

Variable	**n (%)**
City	
Najran City	118 (32.5)
Najran Region	226 (62.2)
Aseer Region	5 (1.5)
Jazan Region	3 (0.8)
Others	11 (3)
Major	
IT	153 (42.1)
Engineering	65(17.9)
Art and Science	49(13.5)
Administrative	92(25.3)
Religion	4(1.1)
Participant has last degree from	

NU	77 (21.2)
King Khaled University KKU	27(7.4)
Jazan University JU	3(0.8)
Others	182(50.1)
Not Applicable	74(20.4)
Have you started a graduate program in the past	
Yes	153 (42.1)
No	210 (57.9)
Do you expect to pursue a graduate degree	
Yes	347 (95.6)
No	16 (4.4)
Ability to communicate in English	
Excellent	67 (18.5)
Very good	103 (28.4)
Good	137 (37.7)
Poor	51 (14)
Very Poor	5 (1.4)

Table III presents the participants motivation to pursue a graduate degree, the expected impact of earning a graduate degree on their careers, and the university they are interested in pursuing their careers on. More than (88%) of the participants are agreeing that that the main motive for them to pursue their career is that because they are passionate about their field of study and they want to learn more. The majority of participants strongly agree that the graduate certificate would improve their career and would increase their chances of getting a new job with (60%) and (58.4%), respectively. When the participants were asked that improving the social status is a primary motive for them to pursue a graduate degree, (48%) strongly agreed and (32%) agreed, while only (9%) did not agree. About (94%) agreed that earning a graduate degree would help them get promotion at their jobs with only (1%) who disagreed. In addition, (62%) agreed that getting a graduate degree would help them changing their current career path. When the participants were asked what do they prefer: a public or private university to pursue their graduate degree at, (98%) were interested in public university, while (27%) were interested in pursuing a graduate degree in private university. When asked about which university they want to pursue their graduate degree at, about (92%) preferred Najran University and that is due to the fact that the majority of the participants (95%) are from Najran region.

TABLE III. PARTICIPANTS' MOTIVATION, EXPECTED IMPACT ON CAREER AND UNIVERSITY INTERESTED IN PURSUING A GRADUATE DEGREE

Item	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
The motivation for pursuing a graduate degree					
Passionate about the field of study	177(48.8%)	147(40%)	29(8.7%)	9(2.5%)	0(0%)
Improving my career	216(60%)	124(34%)	15(4%)	8(2%)	0(0%)
Increasing my chances in getting a new job	212(58.4%)	102(28.1%)	29(8%)	15(4.1%)	5(1.4%)
Improving my social status	176(48%)	116(32%)	39(11%)	24(7%)	8(2%)
The expected impact of earning a graduate degree					
Promotion in my job	244(67%)	98(27%)	17(5%)	3(1%)	0(0%)
Increasing my chances to get a job at better place	209(58%)	106(29%)	35(9.6%)	11(3%)	2(0.4%)
Changing my current career path	135(37%)	89(25%)	54(15%)	71(19%)	14(4%)
Want to pursue your graduate degree in					
Public university	331(91%)	25(7%)	3(1%)	2(0.5%)	2(0.5%)
Private University	46(13%)	52(14%)	92(25%)	118(33%)	55(15%)
Very interested in pursuing my graduate degree in					
Najran University NU	286(79%)	48(13%)	13(3%)	10(3%)	6(2%)
King Khaled University KKU	56(15%)	101(28%)	82(23%)	80(22%)	44(12%)
Jazan University JU	51(14%)	86(24%)	87(24%)	92(25%)	47(13%)
Others	111(30.5%)	54(14.8%)	71(19.5%)	89(24.5%)	39(10.7%)

Table IV presents the majors that the participants would like NU to open a graduate program at. The participants were given a five-point scale to rate their interests in each program. Fourteen majors were presented to the participants, nine of which are IT majors, and 5 are engineering majors. Cybersecurity is the major that got the highest extremely interest with (66.1%) followed by AI and Big Data major with (56.2%). The networks and internet of things (IoT) major came third with (52%). Actually, these three majors are very popular and trendy topics. Also, it is noticeable that the computer science majors are more favored among the participants by getting at least (74.1%) of extremely and very interested votes in every major with exception of game design and development, while engineering majors are getting at their best chance (65.8%) with electrical engineering major. Architectural, civil, chemical and electrical engineering majors have the highest votes on not interested at all option with (5.8%, 5.2%, 4.7%, 4.7%), respectively.

We have already seen from Table II that a significant number of participants want to pursue their graduate degree (95.6%). However, they are lack in their ability to communicate in English where the mode of instruction in IT and engineering field is in English. Only (46.9%) can communicate in English academically. In this study, participants were asked whether they need English as an admission requirement to graduate programs. About (40.30%) participants among an excellent English speaker, about (32%) who are very good at English replied positively. While only (20%) whose English communicative skills are very poor answered positively. The percentage of participants who are already good in English responded positively than that of participants who are weak in English. Overall success score to encompass the English language as graduate program entry requirements using the Chi-square test is considerable with p-value < .05 as presented in Table V. To analyze the result IBM's Statistical Package for the Social Sciences (SPSS) version 22 was used.

TABLE IV. COMPUTER SCIENCE AND ENGINEERING GRADUATE MAJORS THAT PARTICIPANTS WOULD LIKE TO SEE AT NU

Item	**Extremely interested**	**Very Interested**	**Moderately interested**	**Slightly interested**	**Not interested at all**
Computer science	173(47.7)	101(27.8)	44(12.1)	35(9.6)	10(2.8)
Information Systems	173(47.7)	99(27.3)	49(13.4)	32(8.8)	10(2.8)
Software Engineering	182(50.2)	94(25.9)	48(13.2)	29(7.9)	10(2.8)
Cybersecurity	240(66.1)	66(18.2)	33(9)	21(5.8)	3(0.9)
AI & Big Data	204(56.2)	83(22.9)	43(11.8)	28(7.7)	5(1.4)
Networks and IoT	189(52)	106(29.2)	36(9.9)	24(6.7)	8(2.2)
Human Computer Interaction	155(42.7)	114(31.4)	52(14.3)	34(9.4)	8(2.2)
Computer engineering	182(50.2)	101(27.8)	44(12.1)	32(8.8)	4(1.1)
Game Design and development	150(41.3)	80(22.1)	75(20.6)	48(13.2)	10(2.8)
Electrical Engineering	157(43.3)	82(22.5)	64(17.7)	43(11.8)	17(4.7)
Mechanical Engineering	148(40.6)	79(21.6)	75(20.6)	46(12.6)	15(4.6)
Chemical Engineering	142(39.1)	78(21.5)	71(19.7)	55(15)	17(4.7)
Civil Engineering	150(41.3)	79(21.6)	70(19.4)	45(12.5)	19(5.2)
Architectural engineering	141(38.8)	85(23.4)	76(20.9)	40(11.1)	21(5.8)

TABLE V. ASSOCIATION AMONG THE PARTICIPANTS' (RESPONDENTS) ENGLISH LANGUAGE PROFICIENCY AND THEIR WILLINGNESS TO SEE ENGLISH LANGUAGE AS GRADUATE PROGRAMS ENTRY REQUIREMENT USING CHI-SQUARE TEST

Ability to communicate in English	**Participants' opinions and observations over the English Language as the graduate programs entry requirement**			
	Yes **n (%)**	**No** **n (%)**	**Maybe** **n (%)**	***P*-values**
Excellent	27(40.30%)	16(23.88%)	24(35.82%)	X^2(5, *N*=363)=38.97, *P*<0.00001
Very good	33(32.04%)	40(38.83%)	30(29.13%)	
Good	23(16.79%)	73(53.28%)	41(29.93%)	
Poor	4(7.84%)	37(72.55%)	10(19.61%)	
Very Poor	1(20%)	3(60%)	1 (20%)	

TABLE VI. RELATIONSHIP BETWEEN PARTICIPANTS' OCCUPATION (EMPLOYMENT) AND THEIR PREVIOUS ENROLMENT TO GRADUATE PROGRAMS USING CHI-SQUARE TEST

Occupation	**Prior enrolment to graduate programs**		
	Yes **n (%)**	**No** **n (%)**	***P* value**
Government	64(35.96%)	114(64.04%)	X^2(5, *N*=363)=24.60, *P*=0.00006
Private sector	31(41.33%)	44(59.67%)	
Students	5(7.24%)	64(92.76%)	
Unemployed	10(27.78%)	26(72.22%)	
Retired	1(20%)	4(80%)	

Among the participants, about (42.1%) started a graduate program before, according to the data presented in Table II. All the participants in our study belong to a set of occupation having elements include government service holders, private sector, students, unemployment, and retired. According to the research and data revealed from Table VI, it is significant that both government and private sectors employees enrolled in graduate programs previously with (35.96%) and (41.33%) respectively. The overall score is considerable using the Pearson Chi-square test with p-value < .05, as presented in Table VI.

In our study, when asked about the latest degree the participants achieved, only (11%) has completed a graduate degree, including higher diploma, masters, or PhDs according to data presented in Table I. One reason for that is many people start their program; however, they drop out before they finish. One interviewee said he dropped out because it was hard for him to travel every week to Aseer region to attend classes, while some others dropped because they could not study and work at the same time. In our study, a significant number of participants are the employees in both government, and private sectors and still a very considerable number are expecting to pursue their graduate degrees if the afore-mentioned barriers are resolved. Based on that, two more questions were distributed among the participants. One of them is to get their opinions, whether is it feasible to have online classes. The study revealed that among the participants' both government and private sectors employees shown the importance to have online classes with (69.56%). The overall score is considerable using the Chi-square test with p-value < .05, as presented in Table VII. Besides, the second question was to get the participants opinions and responses on the importance of a comfortable schedule and time of classes of graduate programs. The study revealed that among the participants' both government and private sectors employees shown the importance to have a comfortable schedule and time of classes with (95.65%). The overall score is considerable using the Pearson Chi-square test with p-value < .05, as presented in Table VIII.

TABLE VII. RELATIONSHIP BETWEEN PARTICIPANTS' OCCUPATION (EMPLOYMENT) AND THEIR OPINION ON THE IMPORTANCE OF HAVING ONLINE CLASSES USING CHI-SQUARE TEST

Occupation	**Participants' views and observations on the importance to have online classes**			
	Important **n (%)**	**Neutral** **n (%)**	**Not important** **n (%)**	***P* value**
Employee	176(69.56%)	51(20.16%)	26(10.28%)	X^2(4, *N*=363)=35.38, *P*<0.00001
Students	23(33.33%)	24(34.78%)	22(31.89%)	
Unemployed	22(61.11%)	10(27.78%)	4(11.11%)	
Retired	2(40%)	2(40%)	1(20%)	

TABLE VIII. RELATIONSHIP BETWEEN PARTICIPANTS' OCCUPATION (EMPLOYMENT) AND THEIR OPINION ON THE IMPORTANCE OF COMFORTABLE SCHEDULE AND TIME OF CLASSES USING CHI-SQUARE TEST

Occupation	**Participants' views and observations on the importance of a comfortable schedule and time of classes**			
	Important **n (%)**	**Neutral** **n (%)**	**Not important** **n (%)**	***P* value**
Employee	242(95.65%)	6(2.37%)	5(1.98%)	X^2(4, *N*=363)=22.99, *P*=0.000801
Students	60(86.96%)	7(10.14%)	2(2.90%)	
Unemployed	29(80.56%)	5(13.89%)	2(5.55%)	
Retired	3(60%)	1(20%)	1(20%)	

V Discussion and Recommendations

The present research was conducted to analyze the scopes of starting new postgraduate computer and engineering programs at Najran University (NU), Kingdom of Saudi Arabia. In the current study, almost less than half of participants are able to academically communicate in English. Among the participants, about half of responders started a graduate program before, and only 11% was able to complete their study. According to the survey, a significant number of participants want to pursue their graduate degree. However, it is essential to ease the limitations, bridging the barriers to start a new postgraduate computer and engineering programs at Najran University.

It has been found that the native language in Najran is Arabic and mode of instruction in computer and engineering fields is in English. Hence, it is vital to include the English language as one of the graduate programs admission requirements so that students' English proficiency level can be improved. Besides, the dropped out rate is very high among the postgraduate students who get admitted to the programs before. The main reasons for dropped out were the distance the students need to travel every week to Aseer region to attend classes and uncomfortable time table for the students who were doing jobs at the same time. Hence the study revealed that among the participants' both government and private sectors employees shown the importance to have online classes and to have a comfortable schedule and time of courses based on the data presented in Table VII and Table VIII.

VI Conclusion

This study indicated the main barriers and their possible solutions to starting graduate programs at higher education institutions include Najran University (NU), Kingdom of Saudi Arabia. The factors outlined in this study can help higher education institutions and organizations to start graduate programs in finding promising solutions such as increasing English language proficiency, reducing drop rate at graduate levels. An indispensable limitation of the study was recall bias so that some participants might have other reasons for not continuing their previously enrolled graduate programs, but they could not remember it. A widespread future study is recommended, utilizing random sampling and a detailed assessment of responses of respondents.

Acknowledgment

The author would like to express his gratitude to the ministry of education and deanship of scientific research in Najran University, Kingdom of Saudi Arabia for their financial and technical support under code number NU/ESCI/17/077.

References

[1] Saudi Arabia Vision 2030, "Quality of Life Program," Government of Saudi Arabia. Accessed: Jan. 08, 2020. [Online]. Available: https://vision2030.gov.sa/sites/default/files/attachments/QoL%20Arabic_0.pdf,.

[2] Najran Region Manucipality, "Najran Region Manucipality Website." Accessed: Jul. 25, 2020. [Online]. Available: http://www.najran.gov.sa/AboutNajran/Pages/NajranGeographically.aspx.

[3] K. Eriksson, "STARTING POSTGRADUATE EDUCATION AT A SMALL UNIVERSITY," *Öneri Dergisi*, vol. 4, no. 16, pp. 51–54, Jan. 2001, doi: 10.14783/maruoneri.727414.

[4] Cullen, D. J., Pearson, M., Saha, L. J., & Spear, R. H. (1994). Establishing effective PhD supervision. Canberra: AGPS.

[5] Cobbing, S., Maddocks, S., Govender, S., Khan, S., Mbhele, M., Naidoo, K., Tootla, S. & Weston, C., 2017, 'Physiotherapy postgraduate studies in South Africa: Facilitators and barriers', South African Journal of Physiotherapy 73(1), a335. https://doi.org/10.4102/sajp. v73i1.335.

[6] Stathopoulos, I. & Harrison, K., 2003, 'Study at master's level by practising physiotherapists', Physiotherapy 89(3), 158–169. http://dx.doi.org/10.1016/ S0031-9406(05)61032-2.

[7] Gosling, S., 1999, 'Physiotherapy and postgraduate study: A follow-up discussion paper', Physiotherapy 85(3), 117–121. https://doi.org/10.1016/S0031-9406(05)65690-8.

[8] Sran, M.M. & Murphy, S., 2009, 'Postgraduate physiotherapy training: Interest and perceived barriers to participation in a clinical master's degree programme', Physiotherapy Canada 61(4), 234–243. http://dx.doi.org/10.3138/physio.61.4.234.

[9] Glover, P., Bulley, C. & Howden, S., 2008, 'Influences on physiotherapists when deciding to study at Masters level: An exploratory study', Advances in Physiotherapy 10(1), 14–20. https://doi.org/10.1080/14038190701474278.

[10] Murray, R. & Newton, M., 2008, 'Facilitating writing for publication', Physiotherapy 94, 29–34. http://dx.doi.org/10.1016/j.physio.2007.06.004

[11] Hamzat, T.K. & Amusat, N.T., 2002, 'Belief and participation with clinical physiotherapists in research', South African Journal of Physiotherapy 58(2), 32–34.

[12] Huang, R., & Turner, R. (2018). International experience, universities support and graduate employability-perceptions of Chinese international students studying in UK universities. Journal of Education and Work, 31(2), 175–189. https://doi.org/10.1080/13639080.2018.1436751.

[13] Noorelahi, M. M., Soubhanneyaz, A. A., & Kasim, K. A. (2015). Perceptions, barriers, and practices of medical research among students at Taibah College of Medicine, Madinah, Saudi Arabia. Advances in medical education and practice, 6, 479–485. https://doi.org/10.2147/AMEP.S83978.

[14] Al-Zubaidi, K., & Rechards, C. (2010). Arab Postgraduate Students in Malaysia: Identifying and overcoming the cultural and language barriers. Arab World English Journal, 1(1), 107-129.

[15] Sajid, M. R., Laheji, A. F., Abothenain, F., Salam, Y., AlJayar, D., & Obeidat, A. (2016). Can blended learning and the flipped classroom improve student learning and satisfaction in Saudi Arabia?. International journal of medical education, 7, 281–285. https://doi.org/10.5116/ijme.57a7.83d4.

[16] Sandhu, M., Sidique, S., & Riaz, S. (2011). Entrepreneurship barriers and entrepreneurial inclination among Malaysian postgraduate students. International Journal of Entrepreneurial Behaviour and Research, 17(4), 428 - 449. https://doi.org/10.1108/13552551111139656

[17] Jippes, E., Luijk, S.V., Pols, J., Achterkamp, M., Brand, P., & Engelen, J.V. (2012). Facilitators and barriers to a nationwide implementation of competency-based postgraduate medical curricula: A qualitative study. Medical Teacher, 34, e589 - e602.

[18] Symons, M. (2001). Starting a coursework postgraduate degree: The neglected transition. Retrieved July, 8, 2011.

[19] Boyle, Peg & Boice, Bob 1998, 'Best Practices for Enculturation: Collegiality, Mentoring and Structure' in The Experience of Being in Graduate School: An exploration, Ed Anderson, Melissa, New Directions for Higher Education No 101, Jossey Bass Publishers, San Francisco:87-94.

[20] Youngman, Michael 1994, "Supervisors' and Students' Experiences of Supervision' in Postgraduate Education and Training in the Social Sciences: Processed and Products, Ed Burgess, Robert G., Jessical Kingsley Publishers, London & Bristol, Pennsylvania:75-104.

[21] MacKay, Graham 1996, 'Review of the Research Postgraduate Experience at UNE' in Frameworks for Postgraduate Education, Ed Zuber-Skerritt, Outrun, Southern Cross Uni Press, Lismore:126-146.

[22] Golde, Chris 1998, 'Beginning Graduate School: Explaining First-Year Doctoral Attrition' in The Experience of Being in Graduate School: An exploration, Ed Anderson, Melissa, New Directions for Higher Education No 101, Jossey Bass Publishers, San Francisco:55-64.

[23] MacKay, Graham 1996, 'Review of the Research Postgraduate Experience at UNE' in Frameworks for Postgraduate Education, Ed Zuber-Skerritt, Outrun, Southern Cross Uni Press, Lismore:126-146.

[24] Padró, F. F., Erwee, R., Harmes, M. A., Harmes, M. K., & Danaher, P. A. (Eds.). (2018). Postgraduate education in higher education. Springer.

Enhancing Convolutional Neural Network using Hu's Moments

Sanad AbuRass[1], Ammar Huneiti[2], Mohammad Belal Al-Zoubi[3]
Computer Science Department, University of Jordan
Amman, Jordan

Abstract—**Convolutional Neural Networks (CNN) is a powerful deep learning method which is mostly used in image classification and image recognition applications. It has achieved acceptable accuracy in these fields but it still suffers some limitations. One of these limitations of CNN is the lack of ability to be invariant to the input data due to some transformations such as rotation, scaling, skewness, etc. In this paper we present an approach to optimize CNN in order to enhance its performance regarding the invariant limitation by using Hu's moments. The Hu's moments of an image are weighted averages of the image's intensities of the pixels, which produce statistics about the image, and these moments are invariant to image transformations. This means that, even if some changes were made to the image, it will always produce almost the same moments values. The main idea behind the proposed approach is extracting Hu's moments of the image and concatenating them with the flatten vector then feeding the new vector to the fully connected layer. The experimental results show that an acceptable loss, accuracy, precision, recall and F1 score have been achieved on three benchmark datasets which are MNIST hand written digits dataset, MNIST fashion dataset and the CIFAR10 dataset.**

Keywords—CNN; image transformations; invariant; Hu's moments

I. Introduction

Convolutional Neural Networks (CNN) have achieved an acceptable accuracy in classifying images, but it still suffers some limitations [1],[17],[21]. One of these limitations is the lack of ability to be spatially invariant to the input data due to some transformations [14]. Most present approaches usually use dataset augmentation to solve this issue [2],[4],[21], but this needs larger number of model parameters and more training data, and may result in significantly increased training time and larger chance of under- or overfitting [14],[25]. The effect of this issue is even more obvious when dealing with domain-specific problems. E.g. in medical imaging datasets, the rotation can be extraneous due to the symmetric nature of some biological assemblies. However, the scale is constant during imaging process and should not be deemed as a nuisance factor. Moreover, scale-invariance can decrease the performance if object size is informative, for example, in case of classifying healthy cells from cancer cells [15],[28].

Equivariance and invariance are sometimes used interchangeably but these terms are different from each other. "Equivariance" means varying in a similar or "equivalent proportion" while "invariant" means "no variance at all" [6]. More formally, a function f is equivariant with respect to transformation T if $f(T(x)) = T\ (f(x))$. This means that, applying the transformation to x is similarly equivalent to applying the transformation to the result $f(x)$. Invariant is a special case of equivariant. A function f is invariant with respect to a transformation T if $f(T(x)) = f(x)$. this means the result through f does not change when a transformation is applied to the input image [27],[8].

CNN is translation equivariance by nature because of the convolution operations [32], since it convolves all over the input image in order to detect the image's features. So, even if an object was shifted, it will still be detected regardless to its position in the image. Also, pooling operations can make CNN rotation equivariance but only if the object was rotated slightly, but as the degree of rotation increases the CNN may fail to classify the object correctly. Although CNN is translation and slight rotation equivariance, it is not translation, scaling or rotation invariant [5],[7],[24],[26],[32],[31].

The problem of transformation invariant in image classification might cause issues in some fields like robotics and autonomous cars. Because of the movement of the robot or the car, the received images might be distorted, translated, scaled or rotated. Therefore, even if the robot or the car is trained to recognize an object they might fail to do so and might cause problems [17],[3],[35]. Image classification is important in surveillance systems to detect unusual activities. Therefore, the invariant problem might cause problems either by classifying an object to be a threat while it is not then making a false alarm, or by classifying an object as a safe object while it is a threat and, in this case, it might lead to a breach of the system [17]. In health care, image classification is used to classify medical images of the patient in order to help diagnose him/her based on the classified images. The invariant problem in this application might cause issues that lead to a misdiagnose of the patient's condition.

The main objective and contribution of this work is to enhance CNN regarding the invariant limitation in order to achieve higher accuracy in image classification by using Hu's moments of the image [23]. The Hu's moments of an image are weighted averages of the image's intensities of the pixels, which produce statistics about the image, and these moments are invariant to image transformations [18], [36]. This means that, even if some changes were made to the image or if the shape outline got slightly thicker, it will always produce almost the same moments values [9]. Therefore, the Hu's moments of the image can be fed to the CNN in order to make it invariant to image transformations. The taxonomy of this paper will be as follows: Section 2 shows some previous work~~s~~ in this field, Section 3 explains the basics of CNN, Section 4 presents the proposed approach, Section 5 shows the experimental results, and Section 6 shows conclusion.

II. Related Work

Mahesh et al., [18],[19] and Tahmasbi et al. [29] proposed approaches to solve the invariant problem of CNN using Zernike moments. Mahesh et al., [18],[19] proposed a technique which uses Zernike moments in CNN to evaluate the discrimination between face and non-face patterns, and gender classification using facial expression recognition. Their main contribution is the use of Zernike moments as an initial filter, in order to show some unique features of the image that might be helpful to distinguish faces from non-faces image, and gender classifications. They have achieved an accuracy of 100% in distinguishing faces from non-faces images but that is not impressive as it sounds because the discrimination between faces and non-faces is not a hard problem in computer vision any more [10]. In facial expression recognition, they have achieved an accuracy of 87.22%. The main drawback of their work is feature loss. The use of a filter based on Zernike moments might lead to feature loss in some cases.

McNeely-White, et al., [21] Anselmi, et al. [2] and Bruna & Mallat, [4] studied the CNN representations invariance and equivariance to input image transformations. McNeely-White, et al. [21] estimated the linear relationships between representations of the original and transformed images. Although they have achieved good results but their work is considered as data augmentation, and it is not a solid solution to the invariant problem of the CNN.

Cohen & Welling, [7] Gens & Domingos [11] and Mallat [20], analyzed the behavior of the linear representations in relation to symmetry groups, resulting in feature maps that are more invariant to these symmetry groups. Cohen & Welling [7], have revealed that the entire class of such models can be understood mathematically. Although, they have proven their concept mathematically, but their approach still suffers asymmetric world that we live in as described in their own words, “Our approach should also deal better with (approximately) symmetric objects, for which it is not possible to unambiguously estimate pose and motion (what is the pose of a circle?).”. Also, their current model is not suitable for dealing with large images and they consider it as a proof of a concept.

Jaderberg, et al., [14] Hinton [13] and Tieleman [30], have introduced a self-contained module for neural networks. Jaderberg, et al., [14] performed spatial transformations of features by using localization network, parametrized sampling grid, and spatial transformer networks. As in [21] they have used data augmentation which, as it has previously mentioned, is not a robust solution for the invariant problem.

Hinton, et al. [13] and Hinton, et al., [26] proposed a novel CNN architecture which is built up of capsules. These capsules contain group of neurons that are responsible of the instantiation parameters of an entity such as pose velocity and albedo; these capsules will then represent information in a hierarchal from.

The basic theory of their work is that every entity is made up of several smaller entities, so each capsule will try to predict the output of the higher layer capsules, and the capsules which have a greater agreement with the higher layer will be coupled to the parent even more through a positive feedback loop.

Although this work is impressive but it has some shortcomings. The authors have not stated how the weights “W” are learned. Also, the algorithm produces an additional hyper parameter “r” which means more computational complexity. Although the algorithm has achieved the state-of-the-art accuracy on MNIST dataset but it fails to preform so well in CIFAR-10 dataset.

Cheng, et al., [5] Girshick, et al. [12] and Zhang, et al., [34] proposed a method to make CNN rotation invariant. Cheng, et al., [5] added a rotation-invariant layer and Fisher discriminative layer to the CNN in order to make it rotation invariant. These layers will try to learn the objects rotations based on the class, so it can predict the rotation of an object when it recognizes it. They have implemented their algorithm to some famous CNN like VGG and AlexNet, and achieved high accuracy but their work is only directed to rotation invariant, but they did not solve translation or scaling invariant problem in CNN.

Laptev, et al., [15] Su, et al. [28] and Wu, et al., [33] proposed a framework to combine a previous knowledge on nuisance variations with data when training the network. Laptev, et al. [15], formulated a set of transformations and generated multiple images based on these transformations. Then these transformed images are passed through initial layers of the network, and through TI-POOLING operator to from transformation-invariant features. Although they have achieved transformation invariance by pooling transformed features maps, but it added huge computational complexity to the network because of the forward and backward passes for each element.

Worrall, et al., [32] Vedaldi [16] and Memisevic & Hinton, [22] presented a CNN which is equivariant to patch-wise shifting and continuous 360° rotation. Worrall, et al., [32] reconstructed the regular CNN filters by using derivations from complex harmonics, returning a maximal response and orientation for every receptive field patch. Using these derived filters CNN can be invariant to translation and rotation but not scaling. Also, their work has a disadvantage of the higher per-filter computational cost as they must derive and reconstruct all the filters in the CNN.

Up to our knowledge, most of the researches that were studied in the literature review solved the invariant problem of the CNN partially or used data augmentation. In this work, we proposed a general approach to solve the problem with no data augmentation.

III. Convolutional Neural Network

Convolutional Neural Network (CNN) [17] is a major in deep learning which is mostly used in image classification and image recognition tasks due to its convolutional architecture.

Generally, CNN consists of the following phases:

Phase 1: Feature extraction

In this phase, number of filters or kernels will be used to scan the input image, in order to extract features from that image, for example, vertical edges, horizontal edges, corners, etc.

Phase 2: Non-linearity activation

After scanning the filters on the input image, each filter will produce an image which contains the extracted features. The output image must go through a mathematical function which is called Activation Function. In this work, the activation function that will be used is ReLU, which stands for Rectified Linear Unit, which simply converts all the negative values to 0 and keeps the positive values the same as shown in equation (1).

$$R(x) = Max\ (0, x) \tag{1}$$

Phase 3: Pooling

Similar to the Convolutional Layer, the Pooling layer is responsible for reducing the spatial size of the Convolved Feature. This is to decrease the computational power required to process the data through dimensionality reduction.

Phase 4: Dropout

Dropout is used to reduce the CNN overfitting by randomly turning off neurons, so the CNN can take different paths in the training phase. An n-layer fully-connected neural network (ignoring bias) can be defined as:

$$f\,(x;\ \{W_i\}_{i\in\{1,\ldots..,\,n\}}) = \Phi_n\,(W_n\ \Phi_{\,n-1}\,(W_{n-1} \ldots (W_{\,n-1\ldots}\,(\Phi_1\,(W_1 x) \tag{2}$$

Phase 5: Input vector extraction

In this phase, CNN converts the 2D matrix to 1D vector, so it can be fed into the neural network.

Phase 6: Network training using the fully connected Neural Network.

Fully connected layer is a neural network which is used to provide the final classification of an image based on matrix mutilation operations, weights and biases. The input of this phase is the flatten vector which was extracted in the previous phase and the output is the predicted classification. CNN can have several fully connected layers where the output of each layer is the input of the next fully connected layer. The objective of a fully connected layer is to take the results of the convolution/pooling process and use them to classify the image into a label. The output of convolution/pooling is flattened into a single vector of values; each of which represents a probability that a certain feature belongs to a label. For example, if the image is of a cat, features representing things like whiskers or fur should have high probabilities for the label “cat”. Fig. 1 shows an example of how flatten network is fed to the fully connected layer.

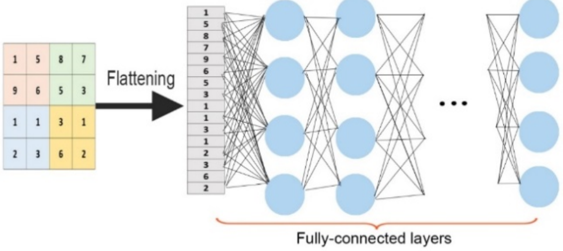


Fig 1. Fully Connected Layer.

IV. Proposed Approach

The main objective of our work is to make CNN invariant to image transformations, in order to achieve higher accuracy in image classification by using Hu’s moments of images [23]. The Hu’s moments of an image are weighted averages of the image’s intensities of the pixels, which produce statistics about the image, and these moments are invariant to image transformations [18], [36]. This means that, even if some changes were made to the image, it will always produce almost the same moments values. Therefore, the Hu’s moments of the image can be fed to the fully connected neural network in order to enhance CNN regarding invariant to image transformations limitation.

The invariant features can be achieved using central moments, which are defined as follows [23], [36]:

$$\mu_{pq} = \iint_{-\infty}^{\infty}(x-\bar{x})^p\ (y-\bar{y})^q\ f(x,y)dxdy \tag{3}$$

Where p,q = 0,1,2,…. , $\bar{x} = \frac{m_{10}}{m_{00}}\ and\ \bar{y} = \frac{m_{01}}{m_{00}}$

The pixel point ($\bar{x}\,,\bar{y}$) are the centroid of the image f (x, y). The centroid moments μ_{pq} computed using the centroid of the image f (x, y) is equivalent to the m_{pq} whose center has been shifted to centroid of the image. Therefore, the central moments are invariant to image translations. Scale invariance can be obtained by normalization [36].

The normalized central moments are defined as follows:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}}\ \gamma = \frac{p+q+2}{2}, p+q = 2,3, \ldots. \tag{4}$$

Based on normalized central moments,[2,3] introduced seven moment invariants:

$$\varnothing_{1\,=}\ \boldsymbol{\eta}_{20} + \boldsymbol{\eta}_{02} \tag{5}$$

$$\varnothing_{2\,=}\ (\boldsymbol{\eta}_{20} - \boldsymbol{\eta}_{02})^2 + 4\boldsymbol{\eta}_{11}^2 \tag{6}$$

$$\varnothing_{3\,=}\ (\boldsymbol{\eta}_{30} - \boldsymbol{3\eta}_{12})^2 + (\boldsymbol{3\eta}_{21} - \boldsymbol{\mu}_{03})^2 \tag{7}$$

$$\varnothing_{4\,=}\ (\boldsymbol{\eta}_{30} + \boldsymbol{\eta}_{12})^2 + (\boldsymbol{\eta}_{21} + \boldsymbol{\mu}_{03})^2 \tag{8}$$

$$\varnothing_{5=}\ (\boldsymbol{\eta}_{30} - \boldsymbol{3\eta}_{12})(\boldsymbol{\eta}_{30} + \boldsymbol{\eta}_{12})[(\boldsymbol{\eta}_{30} + \boldsymbol{\eta}_{12})^2 - 3(\boldsymbol{\eta}_{21} + \boldsymbol{\eta}_{03})^2] + (\boldsymbol{3\eta}_{21} - \boldsymbol{\eta}_{03})(\boldsymbol{\eta}_{21} + \boldsymbol{\eta}_{03})[3(\boldsymbol{\eta}_{30} + \boldsymbol{\eta}_{12})^2 - (\boldsymbol{\eta}_{21} + \boldsymbol{\eta}_{03})^2] \tag{9}$$

$$\varnothing_{6=}\ (\boldsymbol{\eta}_{20} - \boldsymbol{\eta}_{02})[(\boldsymbol{\eta}_{30} + \boldsymbol{\eta}_{12})^2 - (\boldsymbol{\eta}_{21} + \boldsymbol{\eta}_{03})^2\ ] + \boldsymbol{4\eta}_{11}(\boldsymbol{\eta}_{30} + \boldsymbol{\eta}_{12}\ )\ (\boldsymbol{\eta}_{21} + \boldsymbol{\eta}_{03}) \tag{10}$$

$$\emptyset_{7=} (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \quad (11)$$

The adopted research methodology comprises the following steps, as shown in Fig. 2:

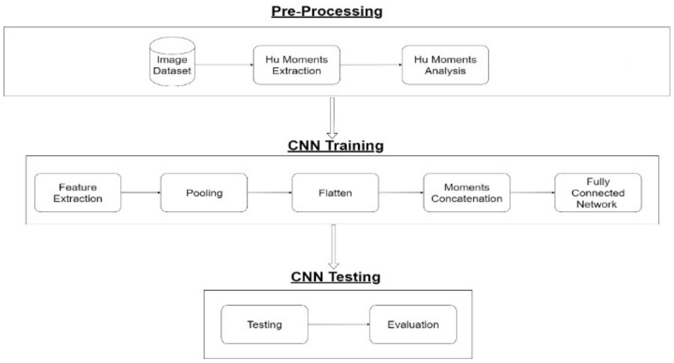


Fig 2. Methodology Phases.

After the flattening operation, Hu's moments of the original image are concatenated with the flattened vector, so the CNN can recognize these values in the testing phase.

Hu's moments concatenation should make the flatten vector more informative and expository. The new vector will be fed to the fully connected network; therefore, CNN will be trained to see the Hu's moments, extent and solidity values alongside with the features vector, these values will affect the neurons' activations in the network in order to achieve transformation invariant. Fig. 3 shows an example of Hu's moments concatenation.

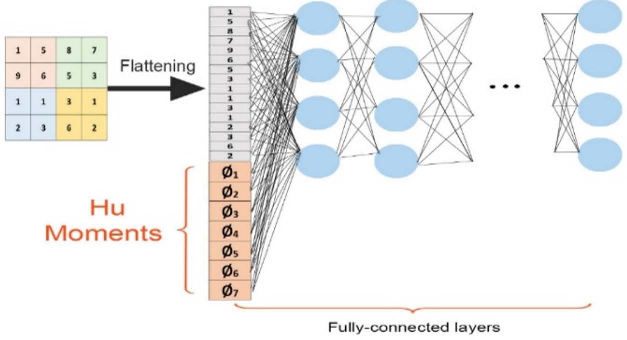


Fig 3. Hu's Moments Concatenation.

V. Experimental Results

We have implemented the proposed approach using Python TensorFlow platform powered by Google Colab notebook. We have tested our approach on three datasets which were MNIST handwritten digits dataset, MNIST fashion dataset and CIFAR10 dataset. Finally, we have compared our results with the work of [18] which uses Zernike moments (ZM) as an initial filter to extract invariant features of the images, as motioned above, by implementing their approach on the three datasets. ZM are projections of an image on to the complex Zernike polynomials that are orthogonal over the unit circle. So, a radius must be provided in order to calculate the ZM of the image. Therefore, we have used the degrees 45° and 90° to extract ZMs of the images. Our approach has archived better loss, accuracy, precision, recall and F1 score compared with the work of [18] on the three dataset MNIST hand written digits, MNIST fashion dataset and CIFAR 10 dataset. The use of Zenick moments as initial filters led to feature loss which led to a decrease in loss, accuracy, precision, recall and F1 score. On the other hand, adding Hu's moments to the flattening vector led to discriminative and more informative vector therefore a better performance.

Table I and Table II shows the results of our approach compared to the results of [18] approach on MNIST handwritten digits dataset. Fig. 4, Fig. 5 and Fig. 6 illustrate the loss, accuracy precision, recall and F1 Score respectively and they show that our approach achieved better performance than [18] approach.

TABLE I. MNIST HANDWRITTEN DIGITS LOSS AND ACCURACY COMPARISONS

Approach	*# of Epochs*	*Loss (cross-entropy)*	*Accuracy*
Original CNN	30	0.062	97.1%
	50	0.022	98.21%
	100	0.018	98.1%
(Mahesh et al. 2017) 45 Degree	30	0.042	97.7%
	50	0.002	98.81%
	100	0.014	98.4%
(Mahesh et al. 2017) 90 Degree	30	0.03	97.8%
	50	0.031	98.1%
	100	0.004	98.4%
Our approach	30	0.004	98.8%
	50	0.006	99%
	100	0.002	99.2%

TABLE II. MNIST HANDWRITTEN DIGITS PRECISION, RECALL AND F1 SCORE COMPARISONS

Approach	*# of Epochs*	*Precision*	*Recall*	*F1 Score*
Original CNN	30	99.22%	99.42%	99.61%
	50	99.72%	99.56%	99.73%
	100	99.81%	99.78%	99.8%
(Mahesh et al. 2017) 45 Degree	30	99.83%	99.72%	99.82%
	50	99.88%	99.77%	99.85%
	100	99.93%	99.82%	99.88%
(Mahesh et al. 2017) 90 Degree	30	99.85%	99.75%	99.84%
	50	99.90%	99.81%	99.87%
	100	99.92%	99.85%	99.89%
Our approach	30	99.94%	99.85%	99.9%
	50	99.95%	99.88%	99.92%
	100	99.96%	99.91%	99.94%

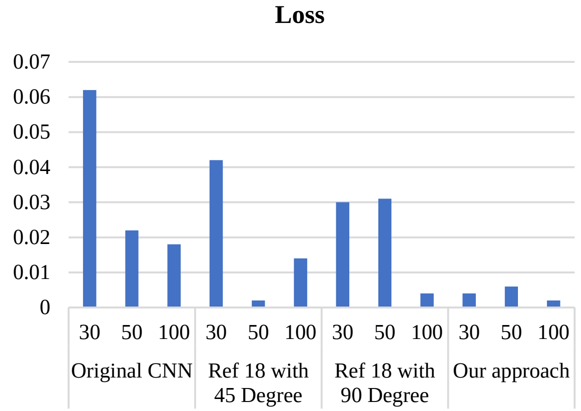


Fig 4. MNIST Handwritten Digits Loss Comparison.

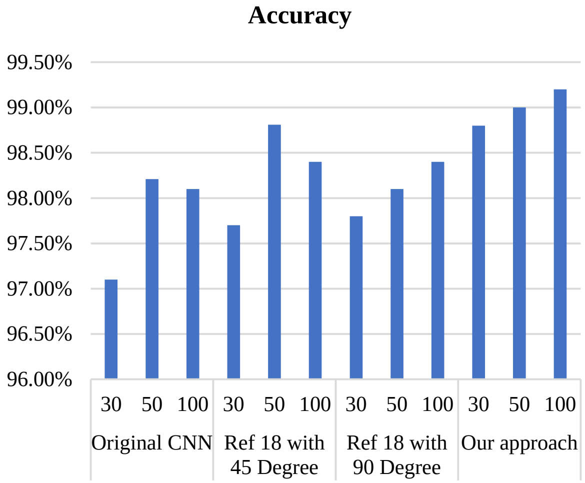


Fig 5. MNIST Handwritten Digits Accuracy Comparison.

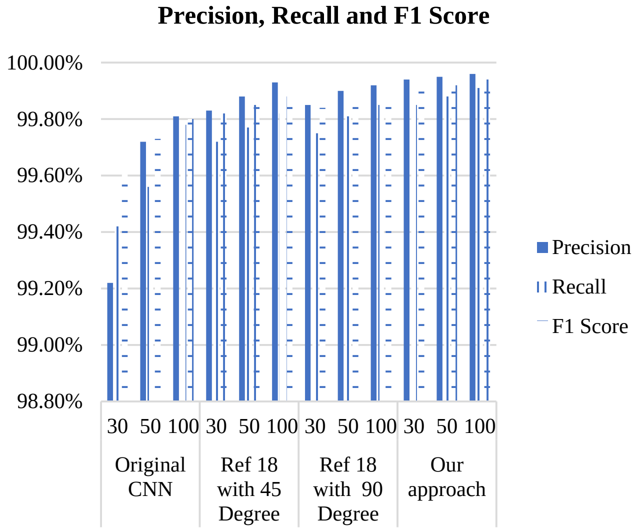


Fig 6. MNIST Handwritten Digits Precision, Recall and F1 Score Comparison.

Table III and Table IV and Fig. 7, 8 and 9 below show the results of our approach implemented on MNIST fashion dataset and we have compared our work with [18] approach and we achieved better results compared to their work.

TABLE III. MNIST Fashion Loss and Accuracy Comparisons

Approach	*# of Epochs*	*Loss (cross-entropy)*	*Accuracy*
Original CNN	30	0.35	83.9%
	50	0.235	84.8%
	100	0.249	86.2%
(Mahesh et al. 2017) 45 Degree	30	0.25	84.6%
	50	0.216	85.9%
	100	0.228	87.1%
(Mahesh et al. 2017) 90 Degree	30	0.268	84.4%
	50	0.224	85.7%
	100	0.176	86.8%
Our approach	30	0.251	89.3%
	50	0.076	90.05%
	100	0.024	91.7%

TABLE IV. MNIST Fashion Precision, Recall and F1 Score Comparisons

Approach	*# of Epochs*	*Precision*	*Recall*	*F1 Score*
Original CNN	30	97.24%	97.9%	97.3%
	50	97.69%	98.04%	97.82%
	100	97.88%	98.19%	98.07%
(Mahesh et al. 2017) 45 Degree	30	97.74%	98.03%	97.88%
	50	97.9%	98.14%	98.02%
	100	97.99%	98.3%	98.15%
(Mahesh et al. 2017) 90 Degree	30	97.82%	97.85%	97.84%
	50	97.94%	98.01%	97.97%
	100	98.01%	98.21%	98.11%
Our approach	30	98.51%	97.85%	98.18%
	50	98.55%	98.14%	98.34%
	100	98.67%	98.28%	98.47%

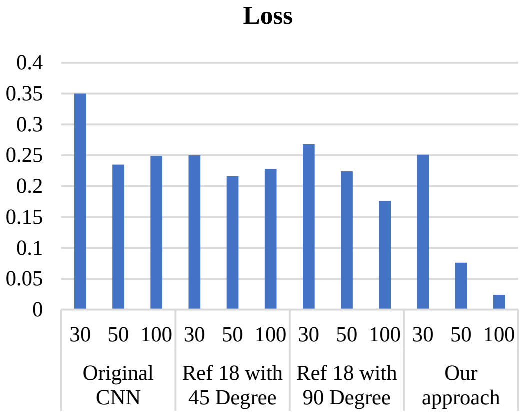


Fig 7. MNIST Fashion Loss Comparison.

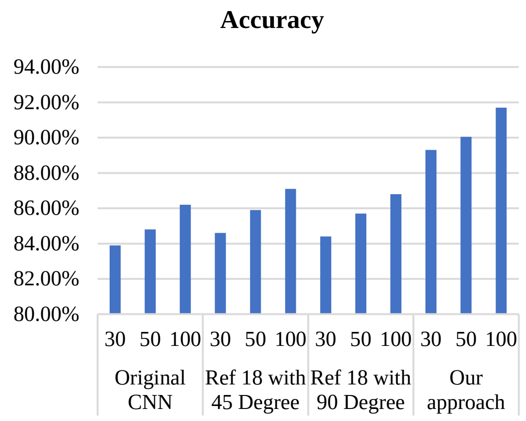


Fig 8. MNIST Fashion Accuracy Comparison.

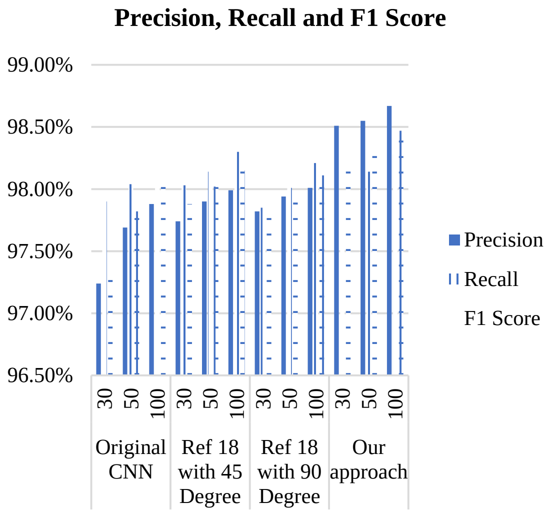


Fig 9. MNIST Fashion Precision, Recall and F1 Score Comparison.

Finally, we have tested our approach and [18] approach on CIFAR10 dataset and we have achieved better performance than their approach as shown below in Table V and Table VI and Fig. 10, 11 and 12.

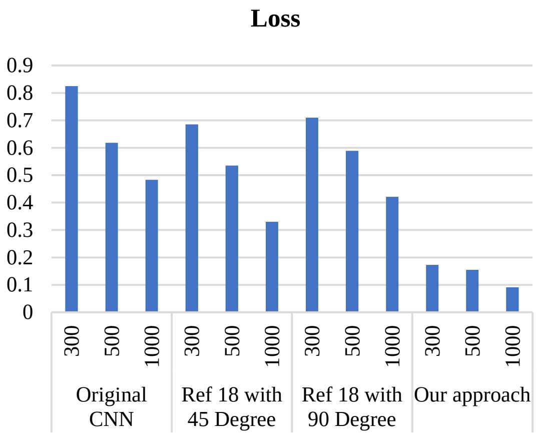


Fig 10. CIFAR10 Loss Comparison.

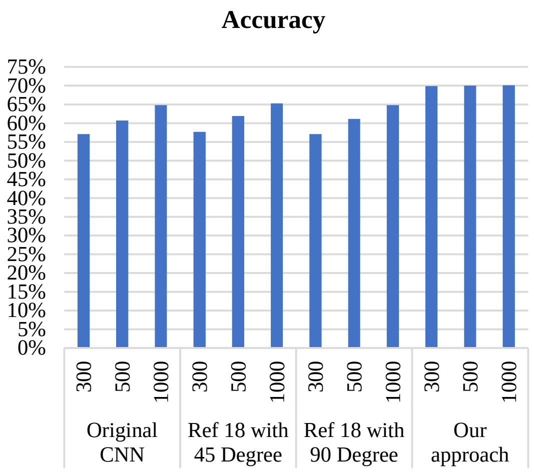


Fig 11. CIFAR10 Accuracy Comparison.

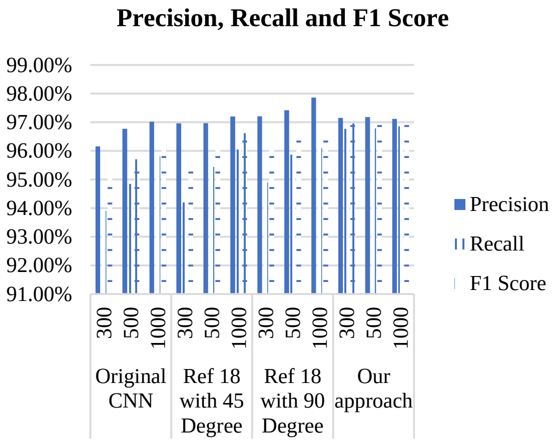


Fig 12. CIFAR10 Precision, Recall and F1 Score Comparison.

TABLE V. CIFAR10 LOSS AND ACCURACY COMPARISONS

Approach	*# of Epochs*	*Loss (cross-entropy)*	*Accuracy*
Original CNN	300	0.825	57.1%
	500	0.618	60.69%
	1000	0.483	64.83%
(Mahesh et al. 2017) 45 Degree	300	0.685	57.7%
	500	0.535	61.9%
	1000	0.33	65.3%
(Mahesh et al. 2017) 90 Degree	300	0.710	57.1%
	500	0.589	61.12%
	1000	0.421	64.8%
Our approach	300	0.173	69.9%
	500	0.155	70%
	1000	0.091	70.1%

TABLE VI. CIFAR10 PRECISION, RECALL AND F1 SCORE COMPARISONS

Approach	*# of Epochs*	*Precision*	*Recall*	*F1 Score*
Original CNN	300	96.16%	93.91%	95.06%
	500	96.77%	94.85%	95.71%
	1000	97.02%	95.81%	96.22%
(Mahesh et al. 2017) 45 Degree	300	96.96%	94.2%	95.56%
	500	96.97%	95.44%	96.2%
	1000	97.2%	96.05%	96.62%
(Mahesh et al. 2017) 90 Degree	300	97.21%	94.9%	96.11%
	500	97.42%	95.87%	96.68%
	1000	97.86%	96.1%	96.78%
Our approach	300	97.15%	96.77%	96.96%
	500	97.18%	96.79%	96.98%
	1000	97.12%	96.86%	96.99%

Fig. 13, 14 and 15 show some real prediction after implementing our approach on MNIST Hand Written Digits Dataset, MNIST Fashion Dataset and CIFAR10 Dataset, respectively.

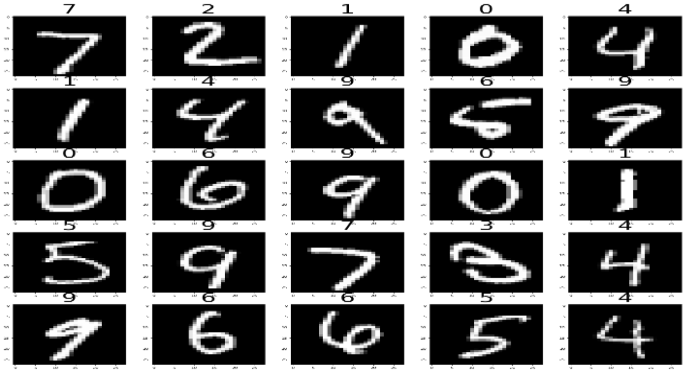


Fig 13. Results of MNIST Hand Written Digits Classification.

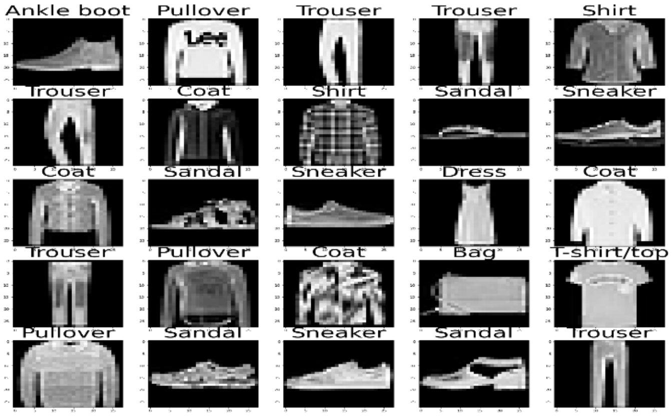


Fig 14. Results of MNIST Fashion Dataset Classification.

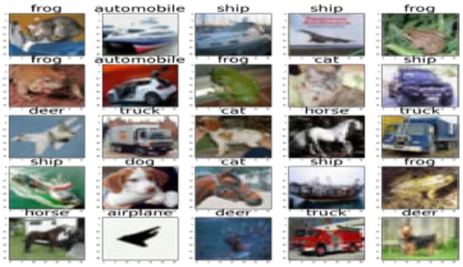


Fig 15. Results of CIFAR10 Dataset Classification.

VI. DISCUSSION

CNN suffers from the problem of being invariant to image transformations. Up to our knowledge most previous researches solved this problem partially or they used data augmentation. Our approach uses Hu's moments to make the flatten vector more descriptive so when it is fed to the fully connected layer it should lead to a better classification regardless to the image transformations. Concatenating the moments with the flatten vector was challenging, since the vector size will increase and it should be the same as the input size of the fully connected layer.

VII. CONCLUSION

This paper presents an approach to enhance CNN regarding the invariant problem by using Hu's moments. The mechanism behind this approach is by concatenating Hu's moments with the flattening vector before feeding it to the fully connected layer in order to make the vector more discriminative and more informative. In this study we have implemented our approach then we have compared out work with the work of Mahesh et al., on the three dataset MNIST hand written digits, MNIST fashion dataset and CIFAR 10 dataset. The results show that our method gave best results in all cases namely loss, accuracy, precision, recall and F1 score. The main limitation of our work is the fixed sizes of the flatten vector that means the size of the vector should precalculated and predefined so it be the same as the size of the input size of the fully connected layer.

REFERENCES

[1] M. Z. Alom et al., "A State-of-the-Art Survey on Deep Learning Theory and Architectures," Electronics, vol. 8, no. 3, p. 292, Mar. 2019, Accessed: Oct. 07, 2020. [Online].

[2] F. Anselmi, J. Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio, "Unsupervised learning of invariant representations," Theor. Comput. Sci., vol. 633, pp. 112–121, Jun. 2016.

[3] M. Billinghurst, A. Clark, and G. Lee, "A survey of augmented reality," 2015, [Online]. Available: http://ir.canterbury.ac.nz/handle/10092/15494.

[4] J. Bruna and S. Mallat, "Invariant scattering convolution networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1872–1886, Aug. 2013.

[5] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning Rotation-Invariant and Fisher Discriminative Convolutional Neural Networks for Object Detection," IEEE Trans. Image Process., vol. 28, no. 1, pp. 265–278, Jan.2019.

[6] B. Chidester, M. N. Do, and J. Ma, "Rotation Equivariance and Invariance in Convolutional Neural Networks," arXiv [stat.ML], May 31, 2018.

[7] T. Cohen and M. Welling, "Group Equivariant Convolutional Networks," in International Conference on Machine Learning, Jun. 2016, pp. 2990–2999, Accessed:Oct.07,2020.[Online].

[8] T. S. Cohen and M. Welling, “Transformation Properties of Learned Visual Representations,” arXiv [cs.LG],Dec.24,2014.

[9] J. Flusser, T. Suk, and B. Zitová, “3D moment invariants to translation, rotation, and scaling,” in 2D and 3D Image Analysis by Moments, Wiley, 2016, p. 96.

[10] K. S. Gautam and T. Senthil Kumar, “Discrimination and Detection of Face and Non-face Using Multilayer Feedforward Perceptron,” in Proceedings of the International Conference on Soft Computing Systems, 2016, pp. 89–103.

[11] R. Gens and P. M. Domingos, “Deep Symmetry Networks,” in Advances in Neural Information Processing Systems 27, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2537–2545.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[13] G. E. Hinton, A. Krizhevsky, and S. D. Wang, “Transforming Auto-Encoders,” in Artificial Neural Networks and Machine Learning – ICANN 2011, 2011, pp. 44–51.

[14] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, “Spatial Transformer Networks,” in Advances in Neural Information Processing Systems 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2017–2025.

[15] D. Laptev, N. Savinov, J. M. Buhmann, and M. Pollefeys, “TI-POOLING: transformation-invariant pooling for feature learning in convolutional neural networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 289–297.

[16] K. Lenc and A. Vedaldi, “Learning Covariant Feature Detectors,” in Computer Vision – ECCV 2016 Workshops, 2016, pp. 100–117.

[17] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, “A survey of deep neural network architectures and their applications,” Neurocomputing, vol. 234, pp. 11–26, Apr. 2017.

[18] V. G. V. Mahesh, A. N. J. Raj, and Z. Fan, “Invariant moments based convolutional neural networks for image analysis,” International Journal of Computational Intelligence Systems, vol.10, no.1, pp. 936–950, 2017.

[19] V. G. V. Mahesh and A. N. J. Raj, “Invariant face recognition using Zernike moments combined with feed forward neural network,” Int. J. Biom., vol. 7, no. 3, pp. 286–307, Jan. 2015.

[20] S. Mallat, “Group Invariant Scattering,” Commun. Pure Appl. Math., vol. 65, no. 10, pp. 1331–1398, Oct. 2012.

[21] D. G. McNeely-White, J. Ross Beveridge, and B. A. Draper, “Inception and ResNet: Same Training, Same Features,” Advances in Intelligent Systems and Computing. pp. 352–357, 2020, doi: 10.1007/978-3-030-25719-4_45.

[22] R. Memisevic and G. E. Hinton, “Learning to represent spatial transformations with factored higher-order Boltzmann machines,” Neural Comput., vol. 22, no. 6, pp. 1473–1492, Jun. 2010.

[23] Ming-Kuei Hu, “Visual pattern recognition by moment invariants,” IRE Transactions on Information Theory, vol.8, no.2, pp.179–187,Feb. 1962.

[24] Y. Poleg, A. Ephrat, S. Peleg, and C. Arora, “Compact CNN for indexing egocentric videos,” in 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Mar. 2016, pp. 1–9.

[25] I. Rocco, R. Arandjelovic, and J. Sivic, “Convolutional neural network architecture for geometric matching,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6148–6157.

[26] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic Routing Between Capsules,” in Advances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 3856–3866.

[27] R. Serfling, “Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardisation,” J. Nonparametr. Stat., vol. 22, no. 7, pp. 915–936, Oct. 2010.

[28] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, “Multi-view convolutional neural networks for 3d shape recognition,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 945–953.

[29] A. Tahmasbi, F. Saki, and S. B. Shokouhi, “Classification of benign and malignant masses based on Zernike moments,” Comput. Biol. Med., vol. 41, no. 8, pp. 726–735, Aug. 2011.

[30] T. Tieleman, Optimizing neural networks that generate images. University of Toronto (Canada), 2014.

[31] S. Vosoughi, P. Vijayaraghavan, and D. Roy, “Tweet2Vec: Learning Tweet Embeddings Using Character-level CNN-LSTM Encoder-Decoder,” in Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, Pisa, Italy, Jul. 2016, pp. 1041–1044, Accessed: Oct. 06, 2020. [Online].

[32] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow, “Harmonic networks: Deep translation and rotation equivariance,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5028–5037.

[33] J. Wu, Y. Yu, C. Huang, and K. Yu, “Deep multiple instance learning for image classification and auto-annotation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3460–3469.

[34] Y. Zhang, K. Sohn, R. Villegas, G. Pan, and H. Lee, “Improving object detection with deep convolutional networks via bayesian optimization and structured prediction,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 249–258.

[35] Z. Zhang and M. Sabuncu, “Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels,” in Advances in Neural Information Processing Systems 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 8778–8788.

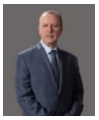
[36] Zhihu Huang and Jinsong Leng, “Analysis of Hu’s moment invariants on image scaling and rotation,” in 2010 2nd International Conference on Computer Engineering and Technology, Apr. 2010, vol. 7, pp. V7–476–V7–480.

[37] Z. Zhu et al., “A Configurable Multi-Precision CNN Computing Framework Based on Single Bit RRAM,” in 2019 56th ACM/IEEE Design Automation Conference (DAC), Jun. 2019, pp. 1–6.

AUTHORS’ PROFILE

Sanad AbuRass received the B.S. degree and M.Sc. degree in computer science from Al Balqa Applied University, As-Salt, Jordan in 2013 and 2016, respectively. From 2015 to 2020 he was ICT Teacher/ IT Administrator at AHSS school, where he worked on school’s management systems. Currently, he is a Ph.D. candidate in computer science at University of Jordan, Amman Jordan. His current interests include image processing, computer vision and deep learning.

Ammar M. Huneiti received his BSc, MSc and PhD degrees from Cardiff University, UK. His BSc is in Computer Science, his MSc is in Information Systems Technologies and his PhD is in Systems Engineering. Between 1992 and 2000 he worked for several private and public sector organizations supervising the design and implementation of IT related projects. In addition, he served as a senior consultant to the ministry of Social Development in Jordan and the National Aid Fund. At present, he is a Full Professor at the Department of Computer Information Systems, King Abdullah II School of Information Technology, the University of Jordan. His research interests include, Intelligent Information Systems, Machine Learning, Spatial Data Mining, and Image Classification.



Mohammad Belal Al-Zoubi is a Professor of Machine Learning and Digital Image Processing in the Department of Computer Information Systems at the Universityof Jordan. He received a B.S. in Cybernetics from The University of Bllgrade in 1985. Prof. Al-Zoubi received his M.S. in Information Systems for the University of Detroit, 1985, and his PhD in Computer Science from Leeds University, UK, 1995. (e-mail: mba@ju.edu.j

Predicting Undergraduate Admission: A Case Study in Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Bangladesh

Md. Protikuzzaman[1], Mrinal Kanti Baowaly[2], Maloy Kumar Devnath[3], Bikash Chandra Singh[4]
Department of Computer Science and Engineering
Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj, Bangladesh[1, 2, 3]
Department of Information and Communication Technology, Islamic University, Kushtia, Bangladesh[4]

***Abstract*—The university admission tests find the applicant's ability to admit to the desired university. Nowadays, there is a huge competition in the university admission tests. The failure in the admission tests makes an examinee depressed. This paper proposes a method that predicts undergraduate admission in universities. It can help students to improve their preparation to get a chance at their desired university. Many factors are responsible for the failure or success in an admission test. Educational data mining helps us to analyze and extract information from these factors. Here, the authors apply three machine learning algorithms XGBoost, LightGBM, and GBM on a collected dataset to estimate the probability of getting admission to the university after attending or before attending the admission test. They also evaluate and compare the performance levels of these three algorithms based on two different evaluation metrics – accuracy and F1 score. Furthermore, the authors explore the important factors which influence predicting undergraduate admission.**

Keywords—Undergraduate admission; educational data mining; XGBoost; Light GBM; GBM; evaluation metrics

I. Introduction

In any country, an undergraduate admission test is one of the most important tests for the students. Students remain conscious about taking admission to their desired universities. In Bangladesh, students who passed the Higher Secondary Certificate (HSC) examination contest the undergraduate admission test. According to the year 2017, 8.01 lakh examinees passed the HSC exam [1] and competed to get admission in different public universities. Universities have their admission requirements for this purpose which are generally based on the students' grade point average (GPA) of the Secondary School Certificate (SSC) and HSC examination, GPA of various courses, etc. However, the total seats in public universities are not sufficient. According to the Ministry of Education of Bangladesh [1], the number of seats in the country's 37 public universities is around 60,000. As a result, about 7 lakhs 40 thousand students did not get the opportunity to study in public universities last year. Even those students who can apply and sit for the admission test do not have the guarantee of admission opportunities in the university because of the limited number of seats. Students have to overcome the barrier of admission test and qualify in the examination to secure their seats. Such students have to go through a long time of mental stress or illness before or after the admission test. The authors realize that this issue cannot be completely removed. But with the aid of modern technologies and strategies e.g. educational data mining, this study can reduce the problem and make students aware of it early in the admission test. If any student can know the pre-examination and post-examination status of a particular university for undergraduate admission, it will be a great benefit for him/her to take the necessary steps to improve the admission test's performance so that he/she can get a chance at the desired university. The authors want to help the students to judge and improve themselves before or after the admission test using this system.

In this study, the authors use the concepts and techniques of data mining which is discovered useful and meaningful information from large-scale data collections [2, 3, 4]. Because of the growing data volume of educational knowledge, educational data mining has a rich area of application [5]. This research is conducted to measure the admission opportunity of a student in Bangabandhu Sheikh Mujibur Rahman Science and Technology University (BSMRSTU), Bangladesh. It is more authentically based on a real dataset collected from the engineering and science faculty students of BSMRSTU. Discovering knowledge from real data gives us a solution that helps students to improve their performance to get admission to BSMRSTU. The authors apply different data mining techniques [6] for a fruitful solution. Here, a total of 500 students' data is collected for this investigation. Though this research focuses on admission chance as the case study, the proposed approach is not restricted to it only. Moreover, this study extensively investigates all possible features or factors of an undergrad candidate and evaluates their impact for predicting admission. The main contributions of this thesis are:

- Developing an admission prediction system for the undergrad students in the engineering faculty at BSMRSTU, Bangladesh.
- Predicting the admission opportunity both before and after the admission test.
- Analyzing and evaluating the possible factors of an admission candidate that affect the admission chance.

The remaining sections of the paper are organized here. In Section II, related works are described. Section III introduces our proposed model for the prediction. The experiment and results of the proposed technique have been described and presented in Section IV. Discussion of the results has been given in Section V. This paper includes the conclusion in Section VI.

II. Related Works

Researchers are working towards the modernization of the education system using education data mining technology. There is a survey [7] paper that depicts the most relevant studies using educational data mining. The researchers concentrate on the field of educational data mining as recent studies show that it is used for analyzing students' performance [8, 9, 10]. There are a number of past studies that focused on predicting admission in colleges or universities. A brief literature review of those studies is presented as follows.

Binu et al. [11] proposed a cloud-based data analysis and prediction system for predicting university admission. There were two modules in the proposed framework, i.e. A Hadoop MapReduce data storage module and an Artificial Neural Network to predict the chances. The data collected had attributes such as status, rank, board, quota, etc. The system did not use academic qualifications in the forecasting process. The neural network had two input nodes, one hidden layer with two nodes, and one output layer with two nodes.

Acharya et al. [12] proposed a comparative approach to predicting graduate admissions by developing four models of machine learning regression: linear regression, vector support, decision tree, and random forest. Roa et al. [13] built a College Admission Predictor in the form of a web application, taking as input the scores obtained by the applicant and his/her personal information and predicting potential college admissions as output.

Ghai [14] developed an American Graduate Admission Prediction model that allows students to choose an apt university by predicting whether or not they will be admitted to the university. Gupta et al. [15] developed a machine learning decision support system for the prediction of graduate admissions in the USA by taking account of certain parameters, including standardized tests, GPA, and Institute Reputation.

Mane and Ghorpade [16] designed a framework for predicting student admission to a particular college using a hybrid combination of Association Rule Mining and Pattern Growth Approach. Data source attributes included student details such as name, gender, caste, address, 10th mark, 12th mark, the score of Common Entry Test, name of pre-college, name of admitted college, and branch. Once valid association rules have been established, the prediction shall be made by the constraint of consequence during the generation of association rules.

Raut and Nichat [17] worked to predict students' performance based on a standard classification methodology, the Decision Tree. This method proposed a model where students take an online test and get an immediate answer (Pass / Fail) coupled with poor principles. The generalization of the sequential pattern mining algorithm was used for the evaluation of output. The decision tree developed by C4.5 is used to assess the success of students and to identify them based on their marks. The author noted that this data mining research could help administrators find poor students and offer extra guidance before the final exam.

Arsad and Buniyamin [18] used the ANN model to forecast the academic success of Bachelor of Technology graduates. The research considered Grade Point (GP) of main subjects that students rate as inputs without taking into consideration their socioeconomic context, thus considering Grade Point Average (GPA) as production. Neural Network (NN) trained engineering graduate students GP to achieve targeted performance. This work showed that core themes have a significant impact on the final CGPA graduation.

Erdogan and Timor [19] used cluster analysis and k-means algorithm techniques to uncover the connection between student entrance test outcomes and their performance. Ktona et al. [20] used the mining association rule as one of the mining partnership tools to classify variables that influence the information acquired by high school students in the ITC course.

Devasia et al. [21] introduced an analysis to predict the success of students in the upcoming academic history test. Build a Web-based program. Nineteen of 700 student characteristics are used as input. When the marks of the pupil were entered, it was contrasted with the scores of the current student, and the ranking of the Naïve Bayes was used to determine the final score. It is noted that the qualification of mother and family income is strongly associated with student success. The collection of data sources, the detection of performance-influencing variables, the construction of a predictive model, and the testing of the model were proposed in the creation of an academic prediction model. The authors noted that this model should help minimize the ratio of loss and help to take appropriate steps against poor performance.

Ruby and David [22] developed a prediction model focused on the Multi-Layer Perceptron algorithm. Datasets were composed of 165 scientific, personal, and economic documents. The overall performance reached for all attributes was 52% and the chosen attributes were 33%.

Aziz et al. [23] created a prediction model that predicts the performance of the first-year computer science students. They used the Naïve Bayes classifier to build their prediction model. By using Naïve Bayes Classifier, it would predict the students' performance level as a categorical value; Poor, Average, and Good. The authors showed that the students' family income, gender, and hometown parameter were the important factors for students' academic performance.

Anuradha and Velmurugan [24] built a new method for predicting the students' final exam results. They applied statistical classification techniques. The experiment shows classifier Naïve Bayes performs better than the other classifiers. The author noted that data mining would improve student status and success at the educational institution.

Kaur et al. [25] used a classification algorithm to classify and viewed slow learners among students using predictive data mining models. From comprehensive literature reviews, variables that affect student success are identified. Both parameters were used as input variables. Five classification algorithms MLP, Naïve Bayes, SMO, J48, and Reptree were applied to the datasets of high school students. MLP was found to have outperformed other classifiers at 75% accuracy. The authors showed that the students who had a computer and internet at home did better during the tests.

In this article, the authors want to determine the chance of a student gets admitted to a university: case study BSMRSTU, Bangladesh. For this, they are using three different machine learning methods that are most effective than others are described. This investigation is different from all the works mentioned above because this study investigates the examinees' possibility of getting a chance in a university before the admission test and after the admission test.

III. Proposed Model

The proposed system overview is displayed in Fig.1. It gives a summary of the possible model of the admission prediction. First, the authors collect data from Bangabandhu Sheikh Mujibur Rahman Science & Technology University (BSMRSTU). After collecting the data, they pre-process the data, extract the features. Then they apply supervised machine learning methods that train, validate the data, and extract knowledge from it. This study predicts the examinee's admission chance before and after the admission to the engineering faculty at BSMRSTU. Each part of this proposed system is described in the following subsections.

A. Data Collection

Having a university admission relies not only on the exam result of the students but also on many other considerations related to their social, economic, cultural, or geographical factors. First, the authors deeply analyze and recognize the causes that are liable or have an impact on admission. To collect data, it is prepared a set of 27 questionnaires as shown in the following subsection *B*. Then the authors provide these questionnaires to the students of BSMRSTU's various departments such as Computer Science and Engineering, Electrical and Electronic Engineering, Electronics and Telecommunication Engineering, Applied Chemistry and Chemical Engineering, Mathematics, Statistics, Chemistry, and Environmental Science & Disaster Management. The first four departments' data is considered for those who got admission in the Engineering faculty (also called A Unit) and the rest four departments' data for those who did not get admission in the Engineering faculty. The total collection is 500 students' data.

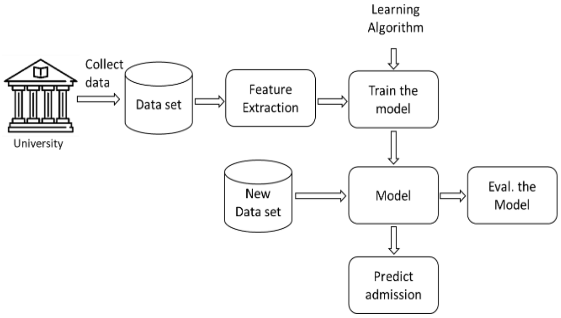


Fig. 1. System Overview.

B. Questionnaires

1. What was your S.S.C GPA (Out of 5.00)?
2. What was your H.S.C GPA (Out of 5.00)?
3. What was your Physics Grade (H.S.C)?
4. What was your Chemistry Grade (H.S.C)?
5. What was your Math Grade (H.S.C)?
6. What was your English Grade (H.S.C)?
7. What was your college?
8. Where did your family live?
   a. Village
   b. Town
9. Name of the Village/Town?
10. District name?
11. About your family education?
    a. Highly Educated
    b. Educated
    c. Less Educated
12. About your family status?
    a. Rich
    b. Middle-Income
    c. Poor
13. Did you live with family at admission time?
    a. Yes
    b. No
14. Where did you live in admission time?
    a. Village
    b. Town
15. Name of the village/town name you lived during admission?
16. District name during admission?
17. Did you get any instruction from any coaching center during admission?
    a. Yes
    b. No
18. Did you have an inspector or motivator?
    a. Yes
    b. No
19. Did you involve in any political party?
    a. Yes
    b. No
20. Did you have any addiction to smoking or any other drug?
    a. Yes
    b. No
21. Did you have an internet facility during admission?
    a. Yes
    b. No
22. What was your daily wasting time (average) on social media or gameplay during admission time (in hours)?
23. What was your daily study time (average) for the admission test (in hours)?
24. Admission test year?
25. Second timer?
    a. Yes
    b. No
26. Expected scores/marks in A Unit-
27. Did you get admission in Unit A?
    a. Yes
    b. No

These features are grouped into two main categories without the target factor. These are (1) before engineering faculty admission (obtained marks in admission test is not included) (2) after engineering faculty admission (obtained marks in admission test is included.

C. Data Preprocessing

The authors prepared the collected data in tabular form from the questionnaire feedback of the students who participated during the data collection. They also applied some data cleaning techniques e.g. handling noise, outliers, missing values, and duplicate data to transform the raw data in a useful and efficient format. The authors considered each questionnaire as a distinct variable or feature for the dataset as shown in Table I. The authors split these 27 features into two categories before and after the admission test takes place to the BSMRSTU Engineering Faculty (Unit A). They allocate the first 25 variables as input features and the last factor (27th variable) as the output label to predict the chance of admission in the engineering faculty before the test. We then allocated the first 26 factors as input features and the last factor (27th variable) as the output label to predict the chance of admission in the engineering faculty after the test happened. The difference between these two categories is of one feature i.e. ‘expected scores/marks in A Unit’ (26th variable in Table I).

For identifying the most relevant input variables (feature selection) to predict undergraduate admission, the authors used the embedded methods which used built-in feature selection methods in machine learning algorithms. All three tree-based machine learning algorithms XGBoost, LightGBM, and GBM used in this investigation have their feature selection method.

D. Description of the Features

A description of the extracted features is given below. Some of the closely related features are discussed together.

S.S.C GPA: S.S.C result is an important factor for identifying a student's quality. It reflects the basic science knowledge of a student. To attend any university admission exam a student should have passed it successfully. So, this factor is important for predicting admission test results. The numeric grade is recorded out of 5.00.

H.S.C GPA: H.S.C result is the most important factor for identifying a student's quality. It also bears the science knowledge of a student. So this attribute is more important for predicting admission test results. The numeric grade is recorded out of 5.00.

Physics grade: A student’s physics grade is the reflection of knowledge on the physics subject. In the BSMRSTU admission test, there are 30 marks on physics for the A unit. So, the grade of physics is more important for getting a chance in the engineering faculty of BSMRSTU. The numeric grade is recorded out of 5.00.

Chemistry grade: Chemistry grade is the reflection of knowledge on the chemistry of a student. In the BSMRSTU admission test of A unit, there are 20 marks on chemistry. So, the grade of chemistry is also important for getting a chance in the engineering faculty of BSMRSTU. The numeric grade is taken out of 5.00.

Math grade: Math grade is the reflection of the knowledge of a student on mathematics. In the admission test of the BSMRSTU A unit, there are 30 marks on math. So, the grade of mathematics plays an important role in getting a chance in the engineering faculty of BSMRSTU. The numeric grade is taken out of 5.00.

English grade: English grade is the reflection of knowledge on the English of a student. We take the numeric grade out of 5.00.

TABLE I. FEATURES LIST

	Feature No.	Features	Value
Before Exam (25 Input Features)	1.	S.S.C GPA	Out of 5.00
	2.	H.S.C GPA	Out of 5.00
	3.	H.S.C Physics Grade	Out of 5.00
	4.	H.S.C Chemistry Grade	Out of 5.00
	5.	H.S.C Math Grade	Out of 5.00
	6.	H.S.C English Grade	Out of 5.00
	7.	College Name	Distinct name
	8.	Family Living Position	Village or Town
	9.	Village/Town name	Distinct name
	10.	District name	Distinct name
	11.	Family Education	3 types
	12.	Family Status	3 types
	13.	Live With Family	Yes or No
	14.	Admission Time Live With Family	Yes or No
	15.	Village/Town Name During Admission	Distinct name
	16.	District Name During Admission	Distinct name
	17.	Coaching Center	Yes or No
	18.	Motivator	Yes or No
	19.	Political Party	Yes or No
	20.	Smoking/Drug	Yes or No
	21.	Internet Facility	Yes or No
	22.	Wasted Time	In Hour
	23.	Reading Time	In Hour
	24.	Admission Test Year	Year
	25.	Second Timer	Yes or No
After Exam(26 Features)	26.	Expected Scores/Marks in Engineering Faculty	Distinct marks
Target Factor	27.	Get admission in Engineering Faculty (Unit A)	Yes or No

College name: The college name is considered to know which college students are willing to admit to the engineering faculty of BSMRSTU. By this feature, we also know the college name that affected most in engineering faculty.

Living area: Bangladesh is a developing country. The development of the country is not equally distributed. The students living in the town area are more concerned about their future rather than the rural area students. Their parents are more conscious of their children's education. The students who live in town have more facilities rather than rural students. This is why it is divided this category into two sectors, living in town and living in the village. Also, including the village or town name and district name which students belong to that town or village.

Family education status: Family education is a great factor for removing the darkness of the mind of a child. Educated parents consider the education of their children as one of the basic needs of life and try hard and soul to provide them with it. They are so much concerned about their children's future. They want to see their children at least as a student in a public university. There are lots of parents who are not educated but they are conscious of their children's future. They also perform a partly role like an educated parent. Family education status also affects admission chances. In a highly educated family parents are more concerned about their child’s education at the early stage. So, the authors divide this category into three sectors- Highly educated, educated, less educated.

Living status during admission: Each student during the admission test either lives in a mess or with a family. The students who live in the mess may have some problems to study. They need to maintain some rules of the mess which kills the time to study. On the other hand, the students who live with family have the extra facility to study and their parents may always take care of them. This is why we divide this attribute into two sectors yes and no. If yes that means, he/she lives with his/her family during admission. If no, then clearly mention the area and district name during admission.

Instruction center: This factor means where a student takes instruction for the admission test. In our country, there are three types of instruction centers to get admissions such as coaching center, batch, and private tuition. In the coaching center, there are lots of students who attend a class. So, the teacher cannot give a student special focus. But in batch or private tuition students can come closer to the teacher and the teacher can focus on each of the student’s preparation. So for the admission test, this factor has special significance. We divide this factor into two categories take coaching or not.

Motivator: The motivator feature plays a vital role in getting a chance in the admission test. Students who do not have any motivator cannot understand what is needed for getting a chance at the university. Even most of them did not know about the university until they admit into a coaching center or batch. So motivator has great importance.

Political involvement: Students who are involved with the political party cannot be able to pay more attention to study because they remain busy with political meetings and processions. We divide this factor into two categories, yes or no.

Frustration and drug addiction: Frustration on anything especially on any matter and drug addiction can hamper a student’s preparation. It also kills time to study. It can reduce the confidence of a student which is must be needed for getting a chance at the university. We divide this factor into two categories, yes or no.

Internet facility: Nowadays the internet is the most important thing or we can say it, teacher, for all types of students in all sectors. So, having an internet connection for collecting previous questions, learning difficult topics and many other purposes is important for a student who wants to get admitted to the university. For this reason, we divide this factor into two categories, yes or no.

Wasted time: Here wasted time means the time a student wastes on social media or playing online games. From many types of research, we know social media has a bad impact on students which hampers the study of a student especially for students who are taking admission tests, it can be like a curse because the students get a small amount of time for admission preparation. So, for this reason, we consider this factor to getting a chance in the admission test. We take this factor individually.

Study time: Study time is the main factor for getting chance in admission test because who studies more time have more opportunity to learn more things. We take this factor individually.

Admission test year: We can see that some years’ questions are easy for getting a chance and some years’ questions are comparatively hard. So the admission test year includes in which year students take part in the admission test.

Second timer: Students who face the admission test the second time have more experience and get more time for taking preparation than the students who face the admission test the first time. So we divide this factor into two categories, yes or no.

Admission test score: An admission test score is most important to predict the examinee's admission after the exam but before the exam, it is not necessary.

Admission test result: This is the main and most important factor which is also called the target factor. Our neural network will predict if a student will get a chance or not. We divide this factor into two categories, yes or no.

IV. Experiment and Results

Since the experimental outcome is very significant in any type of research, all the researchers want to achieve the highest level of accuracy according to their work. This level of accuracy can be different for using different algorithms and methodologies. Researchers must select the algorithm and approach that will provide the highest level of accuracy for the relevant study. In this investigation, the authors predict the admission of the examinees by using different types of supervised learning algorithms. This study applies some advanced algorithms, i.e., XGBoost, LightGBM, Gradient

Boosting Machine to train and validate the predictive model. The authors used k-fold cross-validation (k=5) for validating the model. Therefore, error percentages are lower.

Throughout this experiment, the authors use two specific different measures or metrics to evaluate the quality of classification: accuracy and F1-score based on the following Equations (1) and (4), respectively.

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$\text{Precision}=\frac{TP}{TP+FP} \tag{2}$$

$$\text{Recall}=\frac{TP}{TP+FN} \tag{3}$$

$$\text{F1score}=2*\frac{Precision*Recall}{Precision+Recall} \tag{4}$$

A. Classification Results before the Exam

This study used three machine learning techniques mentioned above to predict the possibility of getting admission in the engineering faculty at BSMRSTU before participating in the admission test. In this case, the obtained marks feature was not in the dataset because this investigation was before the test. The evaluation results to predict admission on test data are summarised in Table II. Note that the accuracy and F1 score are not high, the maximum value is nearly 60 in this case. It is justified as this model is trained and evaluated before the admission test and the authors do not have the admission test score. Nevertheless, the applicant can assess himself to some extent using this model before the admission test.

B. Classification Results after the Exam

This investigation then performs to predict admission opportunities in the engineering faculty at BSMRSTU after participating in the admission test. In this case, expected obtained marks in the exam is used because it is now known to the applicants. The evaluation results are given in Table III. GBM model achieves the highest score 95%. It means the proposed model using the GBM algorithm can accurately predict the admission chance of the applicants after the admission test.

C. Feature Importance before the Exam

The features' importance of the proposed model can be found by using the feature importance property of the model. Feature importance gives a score for each feature of our data, the higher the score is more important or relevant to our output variable. Fig. 2 shows the important features for predicting the admission opportunity of the applicants before the test using the XGBoost learning algorithm. Fig. 3 shows the important features for predicting the admission opportunity of the applicants before the test using the LightGBM learning algorithm. Fig. 4 shows the important features for predicting the admission opportunity of the applicants before the test using the GBM learning algorithm.

The authors combined the feature importance of the aforementioned three learning algorithms. This study did the average feature importance (score) of each feature and plotted in Fig. 5. It shows the features that affect most for getting an opportunity of admitting in engineering faculty before participating in the admission test and these are as follows:

- College name
- Town/Village name
- District name
- H.S.C GPA
- Reading time

TABLE II. CLASSIFICATION RESULTS BEFORE THE ADMISSION TEST

Evaluation Measures	XGBoost	LightGBM	GBM
Accuracy	0.58	0.53	0.50
F1 score	0.56	0.52	0.47

TABLE III. CLASSIFICATION RESULTS AFTER THE ADMISSION TEST

Evaluation Measures	XGBoost	LightGBM	GBM
Accuracy	0.87	0.93	0.95
F1 score	0.84	0.93	0.95

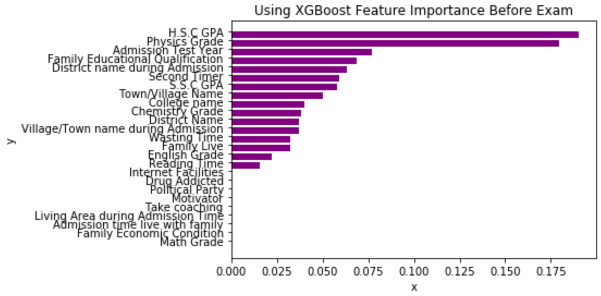


Fig. 2. Feature Importance using XGBoost (before the exam).

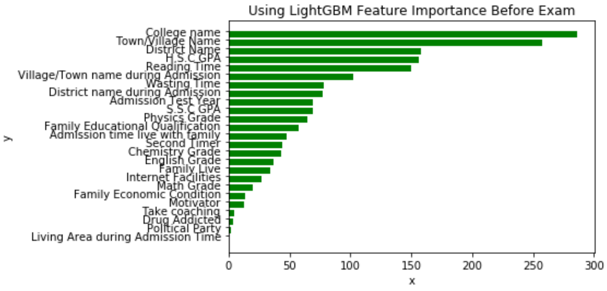


Fig. 3. Feature Importance using LightGBM (before the exam).

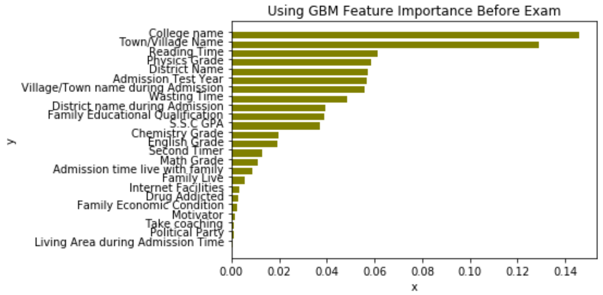


Fig. 4. Feature Importance using GBM (before the exam).

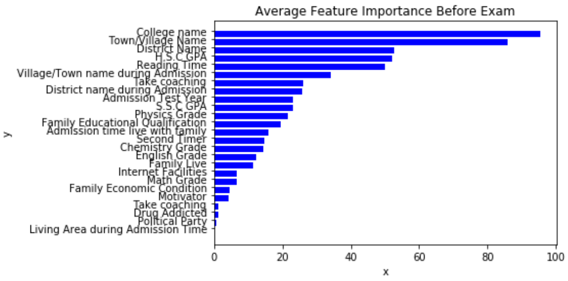


Fig. 5. Average Feature Importance (before the exam).

Here, the academic performance of a candidate e.g. H.S.C GPA and Reading time is an obvious reason for better chances of admission. As we mentioned in Section III, since Bangladesh is a developing country and the development of the country is not equally distributed based on location the address and nativity of the candidate i.e. College name Town/Village name, District name are discovered as important features to predict undergraduate admission before the examination.

D. Feature Importance after the Exam

Fig. 6 shows the important features for predicting the admission opportunity of the applicants after the test using the XGBoost learning algorithm. Fig. 7 shows the important features for predicting the admission opportunity of the applicants after the test using the LightGBM learning algorithm. Fig. 8 shows the important features for predicting the admission opportunity of the applicants after the test using the GBM learning algorithm.

The authors combined the feature importance of the aforementioned three learning algorithms. This study did the average feature importance (score) of each feature and plotted it in Fig. 9. It shows the features that affect most for getting an opportunity of admission in engineering faculty after participating in the admission test and these are as follows:

- Obtained marks
- Admission test year
- Town/Village name
- College name
- H.S.C GPA

Here, the academic performance of a candidate e.g. Obtained marks in the admission test and H.S.C GPA is understandable as the important features to predict undergraduate admission after the examination. Along with these, the address and nativity of the candidate i.e. Town/Village name and College name are discovered as important features because the development of Bangladesh is not equally distributed based on geographical factors. The students who live in town usually have more facilities than rural students. Admission test year is also found as an important feature because it is observed that some years' questions were comparatively easier and some years' questions were comparatively hard which may affect getting a chance in the admission.

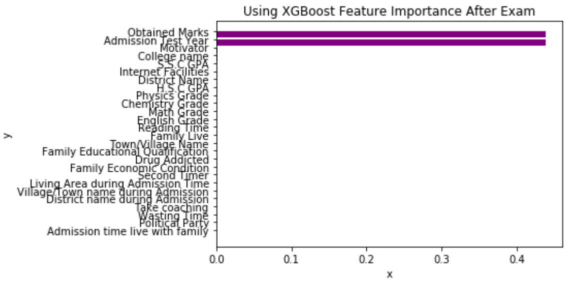


Fig. 6. Feature Importance using XGBoost (after the exam).

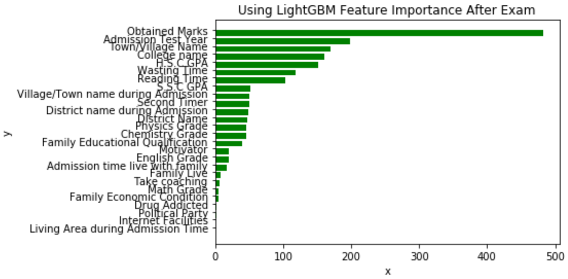


Fig. 7. Feature Importance using LightGBM (after the exam).

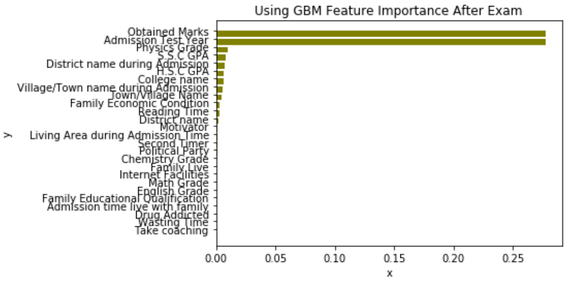


Fig. 8. Feature Importance using GBM (after the exam).

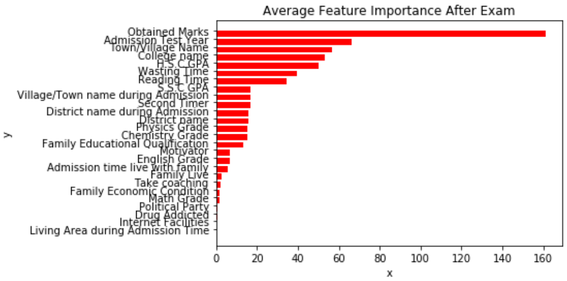


Fig. 9. Average Feature Importance (after the exam).

V. Discussion

The accuracy, F1 score results are different while using different learning algorithms as showed in Table II and Table III. Fig. 10 also shows a comparison among these three algorithms using a bar chart plotting for predicting admission before the test. XGBoost gives the best accuracy and F1 score among the three algorithms although the score is lower as expected. In this case, the dataset did not include the expected score of the applicants in the examination. Fig. 11 shows algorithm comparison for predicting admission after the test. Unlike the first case, the score is achieved higher i.e. above 80%. Here, GBM outperforms XGBoost and LightGBM. The evaluation metrics – accuracy and F1 score both are 95% for GBM. The proposed model using LightGBM and XGBoost also shows excellent results i.e. 93% and above 80% respectively. Hence, the proposed model predicting undergraduate admission after the examination must be very efficient and effective for the students.

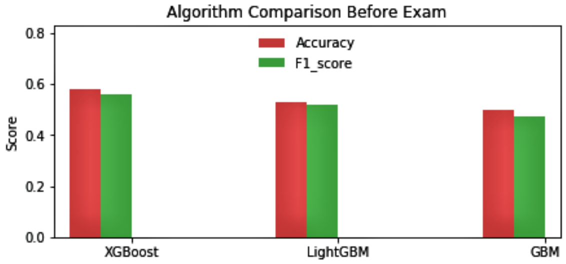


Fig. 10. Algorithm Comparison before the Admission Test.

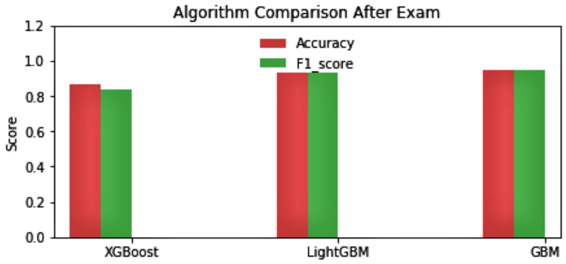


Fig. 11. Algorithm Comparison after the Admission Test.

VI. Conclusion

In this research, the authors used three boosting techniques to estimate the probability of getting an undergraduate admission in the engineering faculty at BSMRSTU, Bangladesh. The root of the dataset is the students of BSMRSTU who are currently studying in the engineering departments and who are not in the engineering departments. Using some machine learning techniques, the authors developed two models separately – the admission predictive model before the admission test and the admission predictive model after the admission test. The authors extensively investigate and analyzed these models. The evaluation results show that the proposed model can able to assist the students in predicting admission opportunities. This study performed the prediction only for the engineering unit at BSMRSTU. This method can be applied to predict admission in any other faculties or universities also.

References

[1] "The Daily Star," 2017. [Online]. Available: https://www.thedailystar.net/backpage/public-universities-admission-still-uphill-battle-1438285. [Accessed 2020].

[2] J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, 3 ed., San Francisco:Morgan Kaufmann Publishers, 2011.

[3] Aggarwal and C. C, Data mining: the textbook, Springer, 2015.

[4] C. C. Aggarwal, Data Mining: The Textbook, 1 ed., Springer International Publishing Switzerland, 2015.

[5] T. Calders and M. Pechenizkiy, "Introduction to the special section on educational data mining," SIGKDD Explorations, vol. 13, no. 2, pp. 3-6, 2012.

[6] J. P. Bigus, Data mining with neural networks: Solving business problems from application development to decision support, McGraw-Hill Companies, 1996.

[7] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 6, pp. 601-618, 2010.

[8] B. K. Baradwaj and S. Pal, "Mining educational data to analyze students' performance," ArXiv Preprint ArXiv:1201.3417, 2012.

[9] El-Halees and A. M, "Mining students data to analyze e-Learning behavior: A Case Study," Mining students data to analyze e-Learning behavior: A Case Study, vol. 29, 2009.

[10] C. Romero, S. Ventura and E. Garcia, "Data mining in course management systems: Moodle case study and tutorial," Computers & Education, vol. 51, no. 1, pp. 368-384, 2008.

[11] P. Binu, A. Chandran and M. Rahul, "A Cloud-Based Data Analysis and Prediction System for University Admission," in 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 2019.

[12] M. S. Acharya, A. Armaan and A. S. Antony, "A comparison of regression models for prediction of graduate admissions," in International Conference on Computational Intelligence in Data Science (ICCIDS), 2019.

[13] A. M. Roa, N. Dharani, A. S. Raghava, J. Buvanambigai and K. Sathish, "College Admission Predictor," Journal of Network Communications and Emerging Technologies (JNCET), vol. 8, no. 4, April 2018.

[14] B. Ghai, "Analysis & Prediction of American Graduate Admissions Process," 2018.

[15] N. Gupta, A. Sawhney and D. Roth, "Will I get in? modeling the graduate admission process for American universities," in IEEE 16th International Conference on Data Mining Workshops (ICDMW), 2016.

[16] R. V. Mane and V. Ghorpade, "Predicting student admission decisions by association rule mining with pattern growth approach," in International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), 2016.

[17] A. B. Raut and A. A. Nichat, "Students Performance Prediction Using Decision Tree Technique," International Journal of Computational Intelligence Research, vol. 13, no. 7, pp. 1735-1741, 2017.

[18] P. M. Arsad and N. Buniyamin, "A neural network students' performance prediction model (NNSPPM)," in International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), IEEE, 2013.

[19] Ş. Z. Erdoğan and M. Timor, "A data mining application in a student database," Journal of aeronautics and space technologies, vol. 2, no. 2, pp. 53-57, 2005.

[20] A. Ktona, D. Xhaja and I. Ninka, "Extracting Relationships between Students' Academic Performance and Their Area of Interest Using Data Mining Techniques," in Sixth International Conference on Computational Intelligence, Communication Systems and Networks, 2014.

[21] T. Devasia, V. T. P and V. Hegde, "Prediction of Students Performance using Educational Data Mining," in International Conference on Data Mining and Advanced Computing, 2016.

[22] J. Ruby and K. David, "Analysis of Influencing Factors in Predicting Students Performance Using MLP-A Comparative Study," International Journal of Innovative Research in Computer and Communication Engineering, vol. 3, no. 2, pp. 1085-1092, 2015.

[23] A. A. Aziz, N. H. Ismail, F. Ahmad and H. Hassan, "A Framework for Students' Academic Performance Analysis using Naïve Bayes Classifier," Jurnal Teknologi, vol. 75, no. 3, 2015.

[24] C. Anuradha and T. Velmurugan, "A Comparative Analysis on the Evaluationof Classification Algorithms in the Prediction of Students Performance," Indian Journal of Science and Technology, vol. 8, no. 15, 2015.

[25] P. Kaur, M. Singh and G. S. Josan, "Classification and prediction based data mining algorithms to predict slow learners in education sector," in International Conference on Recent Trends in Computing, 2015.

Applications of Clustering Techniques in Data Mining: A Comparative Study

Muhammad Faizan[1], Megat F. Zuhairi[2*], Shahrinaz Ismail[3], Sara Sultan[4]
Malaysian Institute of Information Technology, Universiti Kuala Lumpur, Kuala Lumpur, Malaysia[1, 2, 3]
College of Computing and Information Sciences, Karachi Institute of Economics and Technology, Karachi, Pakistan[4]

***Abstract*—In modern scientific research, data analyses are often used as a popular tool across computer science, communication science, and biological science. Clustering plays a significant role in the reference composition of data analysis. Clustering, recognized as an essential issue of unsupervised learning, deals with the segmentation of the data structure in an unknown region and is the basis for further understanding. Among many clustering algorithms, "more than 100 clustering algorithms known" because of its simplicity and rapid convergence, the K-means clustering algorithm is commonly used. This paper explains the different applications, literature, challenges, methodologies, considerations of clustering methods, and related key objectives to implement clustering with big data. Also, presents one of the most common clustering technique for identification of data patterns by performing an analysis of sample data.**

Keywords—Clustering; data analysis; data mining; unsupervised learning; k-mean; algorithms

I. Introduction

Data mining is the latest interdisciplinary field of computational science. Data mining is the process of discovering attractive information from large amounts of data stored either in data warehouses, databases, or other information repositories. It is a process of automatically discovering data pattern from the massive database [1], [2]. Data mining refers to the extraction or "mining" of valuable information from large data volumes [3], [4]. Nowadays, people come across a massive amount of information and store or represent it as datasets[4], [5]. Process discovery is the learning task that works to the construction of process models from event logs of information systems [6]. Fascinating insights, observable behaviours, or high-level information can be extracted from the database by performing data mining and viewed or browsed from various angles. The knowledge discovered can be applied for process control, decision making, information management, and question handling. Decision-makers will make a clear decision using these methods to improve the real problems of this world further. In data mining, many data clustering techniques are used to trace a particular data pattern [2]. Data mining methods for better understanding are shown in Fig. 1.

Clustering techniques are useful meta-learning tools for analyzing the knowledge produced by modern applications. Clustering algorithms are used extensively not only for organizing and categorizing data but also for data modelling and data compression [7]. The purpose of the clustering is to classify the data into groups according to data similarities, traits, characteristics, and behaviours [8]. Data cluster evaluation is an essential activity for finding knowledge and for data mining. The process of clustering is achieved by unsupervised, semi-supervised, or supervised manner [2]. However, there are more than 100 clustering algorithms known and selection from these algorithms for better results is more challenging.

PyClustering is an open-source library for data mining written in Python and C++, providing a wide variety of clustering methods and algorithms, including bio-inspired oscillatory networks. PyClustering focuses primarily on cluster analysis to make it more user friendly and understandable. Many methods and algorithms are in the C++ namespace "ccore::clst" and in the Python module "pyclustering.cluster." Some of the algorithms and their availability in PyClustering module is mentioned in Table I [9].

A. Clustering in Data Mining

Data volumes continue to expand exponentially in various scientific and industrial sectors, and automated categorization techniques have become standard tools for data set exploration [10]. Automatic categorization techniques, traditionally called clustering, helps to reveal a dataset's structure [9]. Clustering is a well-established unsupervised data mining-based method [11], and it deals with the discovery of a structure in unlabeled data collection. The overall process that will be followed when developing an unsupervised learning solution can be summarized in the following chart in Fig. 2:

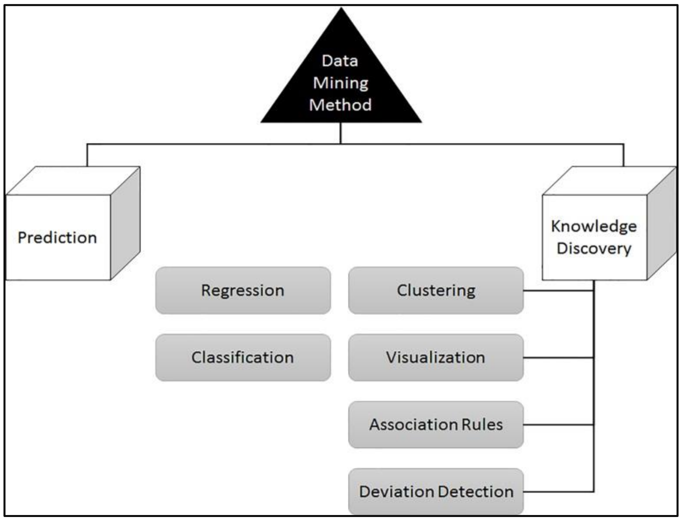


Fig. 1. Methods of Data Mining Techniques.

*Corresponding Author

TABLE I. ALGORITHMS AND METHODS IN "PYTHON MODULE PYCLUSTERING"

Algorithm	Python	C++
Agglomerative (Jain & Dubes, 1988)	✓	✓
BIRCH (Zhang, Ramakrishnan, & Livny, 1996)	✓	
CLARANS (Ng & Han, 2002)	✓	
TTSAS (Theodoridis & Koutroumbas, 2009)	✓	✓
CURE (Guha, Rastogi, & Shim, 1998)	✓	✓
K-Means (Macqueen, 1967)	✓	✓
BANG (Schikuta & Erhart, 1998)	✓	
ROCK (Guha, Rastogi, & Shim, 1999)	✓	✓
K-Medians (Jain & Dubes, 1988)	✓	✓
Elbow (Thorndike, 1953)	✓	✓
GA - Genetic Algorithm (Harvey, Cowgill, & Watson, 1999)	✓	✓
DBSCAN (Ester, Kriegel, Sander, & Xu, 1996)	✓	✓
X-Means (Pelleg & Moore, 2000)	✓	✓
K-Means++ (Arthur & Vassilvitskii, 2007)	✓	✓
Elbow (Thorndike, 1953)	✓	✓
BSAS (Theodoridis & Koutroumbas, 2009)	✓	✓
K-Medoids (Jain & Dubes, 1988)	✓	✓
Sync-SOM	✓	
SyncNet	✓	✓
SOM-SC (Kohonen, 1990)	✓	✓
OPTICS (Ankerst, Breunig, Kriegel, & Sander, 1999)	✓	✓
CLIQUE (Agrawal, Gunopulos, & Raghavan, 2005)	✓	✓
Silhouette (Rousseeuw, 1987)	✓	
MBSAS (Theodoridis & Koutroumbas, 2009)	✓	✓
HSyncNet (Shao, He, Böhm, Yang, & Plant, 2013)	✓	✓
EMA (Gupta & Chen, 2011)	✓	

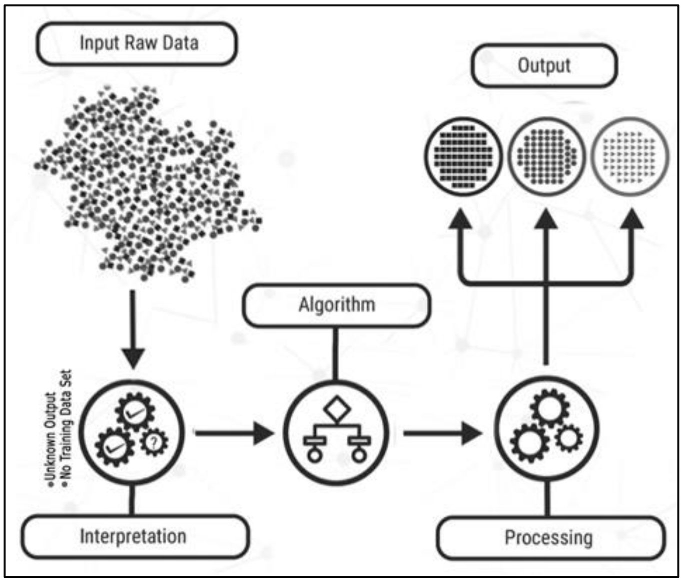


Fig. 2. Unsupervised Learning Model.

The main applications of unsupervised learning are:

- Simplify datasets by aggregating variables with similar attributes.
- Detecting anomalies that do not fit any group.
- Segmenting datasets by some shared attributes.

Clustering results in the reduction of the dimensionality of the data set. The objective of such a clustering algorithm is to identify the distinct groups within the data set [12]. There are different clustering objects, such as hierarchical, partitional, grid, density-based, and model-based [13]. The performance of various methods can differ depending on the type of data used for clustering and the volume of data available [14]. For example, Document clustering has been investigated for use in many different areas of text mining and information retrieval [15]. There are several different metrics of quality, relative ranking, and the performance of different clustering algorithms that can vary considerably depending on which measure is used. Two measures of "goodness" or quality of the cluster are used for clustering. One type of measure allows comparing different cluster sets without external knowledge and is called an "internal quality measure." The other form of measure is called an "external quality measure," which allows evaluating how well the clustering works by comparing the groups generated by the clustering techniques to the classes identified. Fig. 3 shows a simple example of data clustering based on data similarity.

1) Types of clustering: Clustering can generally be broken down into two subgroups:

- Hard Clustering: In hard clustering, each data point is either entirely or not part of a cluster.
  - For example, each customer is grouped into one of 10 groups.
- Soft Clustering: In soft clustering, a probability or likelihood of the data point being in certain clusters is assigned instead of placing each data point into a separate cluster.
  - For example, each customer is assigned a probability to be in 10 clusters.

2) Clustering methodologies: Since the clustering method is subjective, it is the tool that can be used to accomplish plenty of objectives. Every methodology follows several sets of rules and regulations that describe the 'similarity' between data points. Cluster analysis is not an automated task, but an iterative information discovery process or multi-objective collaborative optimization involving trial and error [16]. There are typically more than 100 known clustering algorithms. But few of these algorithms are popularly used. Some of the clustering methodologies are mentioned below in Table II.

The best known and most widely used method of partitioning is K-means [17]–[19]. There are many clustering techniques from which K-means is an unsupervised and iterative data mining approach [11]. The standard approach of all clustering techniques is to classify cluster centres representing each cluster. K-means clustering is a method of cluster analysis aimed at observing and partitioning data point into k clusters in which each observation is part of the nearest mean cluster [7]. The most significant advantage of the K-means algorithm in data mining applications is its efficiency in clustering large data sets. K-means and its different variants have a computation time complexity that is linear in the number of records but is assumed to discover inferior clusters [15].

The K-means algorithm is a basic algorithm for iterative clustering. It calculates the distance means, giving the initial centroid, with each class represented by the centroid, using the distance as the metric and given the classes K in the data set. In the k-means partitioning algorithm, the mean value of objects within-cluster is represented at the centre of each cluster.

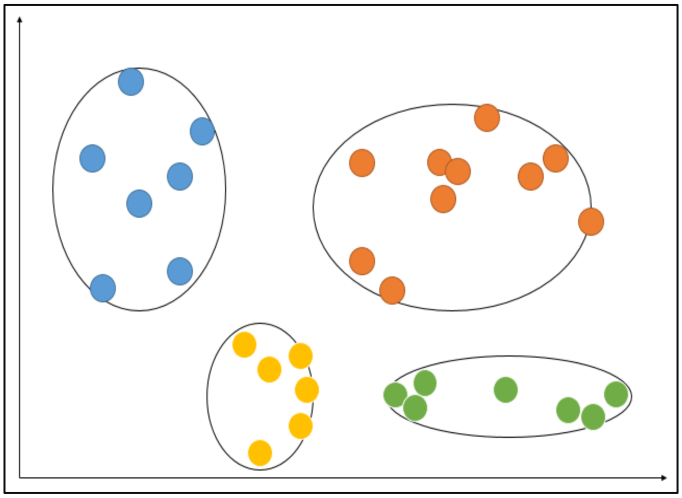


Fig. 3. Simple Clustering Example.

TABLE II. CLUSTERING METHODOLOGIES

Typical Clustering Methodologies	
Method	***Algorithm***
Distance-based method	• Partitioning algorithms "K-means, K-medians, K-medoids." • Hierarchical algorithms, "Agglomerative, Divisive method." These algorithms run iteratively to find the local optima and are incredibly easy to understand but have no scalability for handling large datasets.
Grid-based method	• Grid-base algorithm: Individual regions of the data space are formed into a grid-like structure. These methods use a single-uniform grid mesh to separate the entire problem domain into cells. The cell represents the data objects located within a cell using a collection of statistical attributes from the objects.
Density-based method	• Density-Based Spatial Clustering of Applications with Noise / DBSCAN • Ordering points to identify the clustering structure OPTICS These algorithms scan the data space for areas with different data points density within the data space. It isolates different density regions within the same cluster and assigns the data points within those regions.
Probabilistic and generative models	• Expectation-maximization algorithm: Modeling data from a generative process. Often these models suffer from over-fitting. A prominent example of such models is the Expectation-Maximization algorithm that uses normal multivariate distributions.

II. Background and Discussion of Clustering Applications and Approaches

Cluster analyses have lots of applications in different domains, e.g., It has been popularly used as a preprocessing step or intermediate step for other data mining tasks "Generating a compact summary of data for classification, pattern discovery, hypothesis generation and testing, compression, reduction, and outlier detection, etc." Clustering analysis can also be used in collaborative filtering, recommendation systems, customer segmentation, multimedia data analyses, biological data analyses, social network analysis, and dynamic trend detection. Some of the clustering techniques and approaches are discussed in Table III.

A. Requirement and Challenges

Despite recent efforts, the challenge of clustering on "mixed and categorical" data in the sense of big data remains, due to the lack of inherently meaningful similarity measurement between the high computational complexity of current clustering techniques and categorical objects [18]. For cluster analysis, there are several items to be considered. Some of them are mentioned in Table IV.

Typically, there are multiple ways to use or apply clustering analysis; some advantages and limitations of clustering techniques are mentioned in Table V.

- As a stand-alone tool to get insights into data distribution.
- As a preprocessing (or intermediate) steps for other algorithms.

According to [24], [25], parallel classification is a better approach for big data, but due to its implementation's complexity remains a significant challenge. However, the framework of MapReduce can be suitable for implementing parallel algorithms, but still, there is no algorithm to handle all Challenges of big data. In [26], the authors proposed a novel Spark extreme learning machine "SELM" algorithm based on a spark parallel framework to boost the speed and enhance the efficiency of the whole process. SELM gives the highest speed and minimal error in all experimental results compared to Parallel Extreme Learning Machine (PELM) and an improved Extreme Learning Machine (ELM*). Table VI presents the pros and cons of different clustering algorithms with real-world applications.

TABLE III. Clustering Techniques and Approaches with Benefits

Ref.	**Author**	**Year**	**Technique / Algorithm**	**Approach**	**Outcomes**
[19]	Chunhui Yuan and Haitao Yang	2019	K-Means Clustering Algorithm	Different methods applied to each dataset to determine the optimal selection of K-Value.	Concluded that these four methods (Elbow methods, silhouette coefficient, gap statistics, and canopy) satisfy the criteria for clustering small data sets. In contrast, the canopy algorithm is also the best choice for large and complex data sets.
[20]	Tengfei Zhang, Fumin Ma	2015	Rough k-means clustering	Improved rough k-means clustering with Gaussian function based on a weighted distance measure	An improved rough k means algorithm based on weighted distance measure with Gaussian function handles the objects which are wrongly assigned to clutters, also handles vulnerable sets while distributing *overlapping* objects in different clusters by rough k means with the same weighted distance for both upper and lower bounds.
[21]	Lior Rokach, Oded Maimon	2015	Clustering methods: Hierarchical- based, Model-based, Grid-based, Partitioning based, Density-based.	Different clustering methods/techniques are used to determine clustering efficiency in large data sets and explain how the number of clusters can be calculated.	For large dataset concluded that "K-means clustering is more efficient in terms of its time, space complexity, and its order-independent" and "Hierarchical clustering is more versatile, but it has the following disadvantages: Time complexity O($m^2 * logm$) and space complexity of a hierarchical agglomerative algorithm is O(m^2).
[22]	Zengyou He, Xiaofei Xu, Shengchun Deng, Bin Dong	2015	K-mean, K-modes, K-Histogram	Compare different clustering algorithms to determine an efficient clustering algorithm for the categorical dataset.	K-Histogram is the enhanced version of K-means to categorical areas by substituting means of clusters with histograms. In general, K-Histogram is almost similar to the K-modes algorithm, but as compared to k-modes, k-histogram algorithms are more stable, and the algorithm will converge faster.
[16]	M.Venkat Reddy, M. Vivekananda, RUVN Satish.	2017	Divisive, and Agglomerative Hierarchical Clustering with K-means.	Discover an efficient clustering by comparing Divisive and Agglomerative Hierarchical Clustering with K-means.	To obtain high accuracy, Agglomerative Clustering with k-means will be the practical choice. Divisive clustering with K-means also works efficiently where each cluster can be taken fixedly.
[23]	Ahamed Al Malki, Mohamed M. Rizk, M.A. El-Shorbagy, A. A. Mousa	2016	K-means, Genetic algorithm	For solving the clustering problems, introduced a hybrid approach of the Genetic algorithm with K-means.	A hybrid approach of K-means with a Genetic algorithm efficiently solves all the problems of the k-means, e.g., K-mean will produce empty clusters with initial centre vector and converge to non-optimal value, etc.

[7]	Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta.	2012	Hierarchical, K-Means, DB Scan, OPTICS, Density-Based Clustering, EM Algorithm	A comparison was made between different clustering techniques to measure the best performing algorithm.	K-means is faster than all the algorithms that are discussed in this paper. When using a huge dataset, K-means and EM will the best results than hierarchical clustering.
[11]	Karthikeyan B., Dipu Jo George, G. Manikandan, Tony Thomas.	2020	K-means, Agglomerative Hierarchical Clustering	Comparative research to determine the best-suited algorithm on K-Means and Agglomerative Hierarchical Clustering.	The k-means is best suited for larger datasets in term of minimum execution time and rate of change in usage of memory. It is also concluded that agglomerative clustering is best suited for smaller datasets due to the overall minimum consumption of memory.

TABLE IV. CONSIDERATIONS OF CLUSTERING ANALYSIS

Considerations for Clustering Analysis		
Considerations	***Options***	***Examples***
Similarity measure	Distances-based / Connectivity- based	Euclidean, road network, vector / Density, Contiguity
Partitioning criteria	Single level / Hierarchical partitioning	Often / Multi-level
Cluster space	Full space / Subspace	Low-dimensional / High-dimensional
Separation of clusters	Exclusive / Non-exclusive	Datapoint belongs to only region / Data point belongs to multiple regions

TABLE V. ADVANTAGES AND LIMITATIONS OF CLUSTERING TECHNIQUES

Clustering techniques “Advantages & Limitations”		
Clustering Techniques	***Advantages***	***Limitations***
Data-mining clustering algorithms	• Implementation is simple. • Compromises on user’s privacy.	• Do not deal with a large amount of data
Dimension reduction	• It is very fast, reduces the dataset, and the cost of the treatment will be optimized.	• It must be applied before the classification algorithm. • It cannot provide an efficient result for the high dimensional dataset. • It may lose some amount of data.
Parallel classification	• It gives minimal execution time and more scalable.	• Difficult to implement.
MapReduce framework	• Flexibility, scalability, security and authentication, batch processing, etc.	• It does not do best for graphs, iterative, and incremental, multiple inputs, etc.

TABLE VI. CLUSTERING ALGORITHMS PROS AND CONS

Algorithm Name	**Pros**	**Cons**	**Applications in Real World**
K-means	➢ Handles large amounts of data. ➢ Minimum execution time. ➢ Simple to implement, etc.	➢ Manually choose the K value. ➢ Clustering outliers. ➢ Dependent on starting point/value. ➢ Handle empty clusters, etc.	➢ Wireless networks. ➢ System diagnostic. ➢ Search Engine. ➢ Document Analysis. ➢ Fraud detection. ➢ Call record Analysis.
Hierarchical Clustering	➢ Do not need to specify the initial value. ➢ Easy to implement, scalable and easy to understand, etc.	➢ Cannot handle a large amount of data with different sizes. ➢ No backtracking. ➢ No swapping between objects. ➢ More space and time complexity.	➢ Humans skin analysis [27] ➢ Generating a portal site. ➢ Web usage mining.
Genetic Algorithm	➢ Easily understandable and converge with different problems. ➢ It cannot always give the best result for all problems but provide the optimum solution. ➢ It cannot search for a single point, search from a population of a point.	➢ It is computationally expensive, e.g., time-consuming. ➢ It may lose data in a crossover.	➢ Engineering Designs. ➢ Robotics. ➢ Telecommunications, traffic, shipments routing. ➢ Virtual Gaming. ➢ Marketing.
DBSCAN	➢ The number of clusters does not need to be defined. ➢ Handle outliers.	➢ Unable to handle datasets with distinct densities. ➢ Struggles to work with High Dimensionality Data.	➢ Satellite pictures, etc.

III. Running Example with K-mean

A car manufacturer company wants to identify the purchase behaviours of its customers to view which product is getting more sales and what is the procedure of our customers. They are currently looking at each customer's details based on this information, decide which product manufacturing should be increased and what are the behaviour of customers which helps the company to monitor sales for other products by starting a promotional campaign or increase the availability of resources.

Recently, the company can potentially have millions of customers. It is not possible to look at each customer's data individually and then make a decision. A manual process will take a huge amount of time. This is when K-means Clustering assists in a convenient way to analyze data automatically. The K-mean clustering algorithm utilizes a fixed number of clusters for optimum clustering [12], [28]. Initially, start partitioning with the chosen number of clusters next to improve the partitions iteratively to find the patterns in data. Let D= {D1, D2, …, Dn} be the set of data points and Y= {Y1, Y2, …, Yt} be the set of centers. This clustering technique is implemented and analyzed using a k-mean clustering tool WEKA. In the following steps, the K-means algorithm can be implemented:

K-mean Clustering Algorithm

1. The first step in k-means is to pick the number of clusters, k.
2. Randomly select k number of clusters centre.
3. Find the distance between each data point and each cluster centre.
4. In contrast with other cluster centres, assign the data point to the nearest cluster centre.
5. Discover the new Cluster Centre again $Y_i = \sum_{i=0}^{ki} D_i$ by where k_i represents number of data points in i^{th} cluster.
6. Once again, find out the distance between each data point and the new cluster centre.
7. If no data points were reassigned then stop, otherwise back to step 3.

The data set used for the K-mean clustering example will focus on a fictional car dealership. The dealership is starting a promotional campaign for slow-selling units, whereby it is trying to push resources to its valuable customers. Table VII shows the sample dataset, which is used for the analysis.

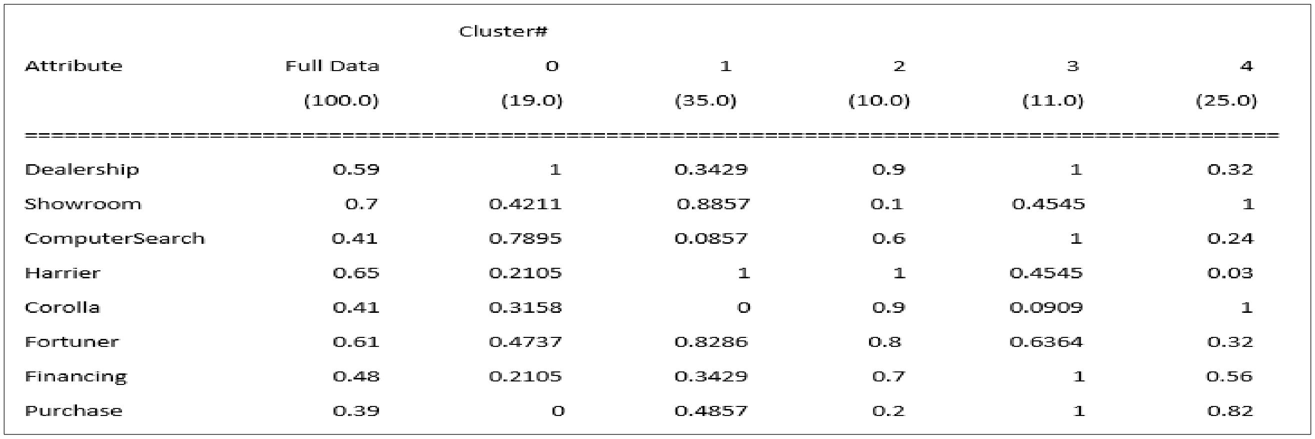
In Table VII, every row shows the purchase behaviour of customers, e.g., Customers went to the dealership without going on a showroom and done some computer search mostly interested in Toyota Harrier without financing they purchased it. These types of behaviour understandings about customers help Toyota to manage their sales. K-mean clustering allows the company to perform analysis without any efforts by finding patterns in a given dataset, shown in Fig. 4 and Fig. 5.

Fig. 4 explains that based on cluster 3, 100% of customers went for the dealership, whereas 45% went to the showroom too, and 100% of the customers also did computer searching. The majority of the customer that is 63% have shown interest in Fortuner, whereas 45% had shown interest in Harrier, and the least interest was found to be 9% in Corolla. These customers who 100% end up financing and purchasing a product consistently went to the dealership and done computer searching before buying an SUV car.

Meanwhile, based on cluster 4, only 32% of customers went to the dealership, whereas 100% went to the showroom, and 24% also did computer searching. Majority of the customers that are 100% interested in Corolla whereas 32% had shown interest in Fortuner, and the least interest was found to be 3% in Harrier; out of all these, 56% of the customers went for the financing details whereas 82% ends up purchasing a product. These are the customers looking for a small family car, i.e., Corolla, mostly approaching the showrooms.

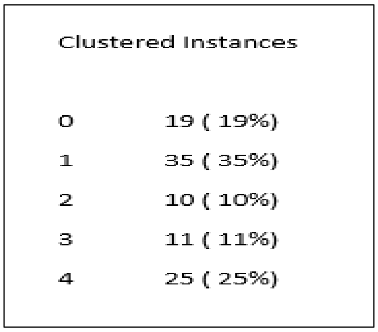
TABLE VII. Sample Dataset

No.	Dealership	Showroom	Computer Search	Harrier	Corolla	Fortuner	Financing
1	1	0	1	0	1	0	0
2	0	1	0	0	0	1	0
3	1	0	0	1	0	0	1
4	1	0	1	1	0	0	1
5	0	1	1	0	1	1	1
6	0	0	1	0	1	0	0
7	0	0	1	1	0	1	1
8	1	1	0	0	0	1	0



```
                                   Cluster#
Attribute              Full Data          0          1          2          3          4
                         (100.0)     (19.0)     (35.0)     (10.0)     (11.0)     (25.0)
=====================================================================================
Dealership                  0.59          1     0.3429        0.9          1       0.32
Showroom                     0.7     0.4211     0.8857        0.1     0.4545          1
ComputerSearch              0.41     0.7895     0.0857        0.6          1       0.24
Harrier                     0.65     0.2105          1          1     0.4545       0.03
Corolla                     0.41     0.3158          0        0.9     0.0909          1
Fortuner                    0.61     0.4737     0.8286        0.8     0.6364       0.32
Financing                   0.48     0.2105     0.3429        0.7          1       0.56
Purchase                    0.39          0     0.4857        0.2          1       0.82
```

Fig. 4. Customers Purchase Behaviours.



```
Clustered Instances

0       19 ( 19%)
1       35 ( 35%)
2       10 ( 10%)
3       11 ( 11%)
4       25 ( 25%)
```

Fig. 5. Clustered Instances based on Customers Behaviour.

IV. Conclusion

This paper describes the different algorithms and methodologies used to handle large and small sets of data. The process of clustering is to group data based on their characteristics and similarities. Previously described the clustering models, many clustering techniques used to partition the data into a set of clusters. Algorithm selection should depend on the properties and the nature of the data collection because each algorithm has its pros and cons. This shows that there is no algorithm to manage all the clustering challenge. However, there are some algorithms to provide an optimist solution based on their sufficiency to face the challenges of the problem. To achieve high accuracy in terms of time and space, K-means would be the best choice for large and categorical data. However, we need to reduce their time and memory's complexity by upgrading Clustering Algorithms. However, a combined approach of the Genetic Algorithm with K-means can almost resolve all the issues of K-means. Genetic K-means Algorithm (GKA) speeds up the convergence to a globally optimum, and it concludes that GKA is faster than evolutionary Algorithms.

V. Future Directions and Open Issues

To date, Data Mining and information disclosure are advancing an essential innovation for businesses and scientists in numerous domains. Although information mining is extremely powerful, it faces innumerable difficulties during its usage. The problems could be identified with performance, data, strategies, and procedures utilized. The information mining measure becomes effective when the challenges or issues are distinguished accurately and sifted through appropriately.

Some of the following challenges and future directions are:

- Efficiency and Scalability of Algorithms: The data mining algorithms must be proficient and adaptable to extricate data from gigantic sums of information within the database. So, as a future direction, develop a parallel formulation of an Improved rough k-means algorithm to enhance the efficiency of an algorithm.
- Privacy and Security: Information mining ordinarily leads to genuine issues in terms of information security, protection, and administration. For case, when a retailer reveals his clients purchasing details without their permission. So, as a future direction, there needs to develop a single cache system and DES (Data Encryption Standard) techniques in any Clustering Algorithm to improve the privacy and security of data in the cloud.
- Complex Data Types: Complex data elements, objects with graphical data, temporal data, and spatial data may be included in the database. Mining of these types of data isn't practical to be done one device.
- Performance: The execution of the data mining framework depends on the proficiency of calculations and procedures are utilizing. The calculations and strategies planned are not up to the marked lead to influence the performance of the data mining process.

Therefore, as a future direction, we need to introduce a new hybrid approach of an Improved Rough k-means Algorithm, and the Genetic Algorithm will improve the performance and handles the complex data. The combination of Partitioning Clustering and Hierarchical Clustering Algorithms will also increase the accuracy of data analysis.

References

[1] S. Sharma, J. Agrawal, S. Agarwal, and S. Sharma, "Machine learning techniques for data mining: A survey," 2013 IEEE Int. Conf. Comput. Intell. Comput. Res. IEEE ICCIC 2013, no. I, 2013.

[2] M. Z. Hossain, M. N. Akhtar, R. B. Ahmad, and M. Rahman, “A dynamic K-means clustering for data mining,” Indones. J. Electr. Eng. Comput. Sci., vol. 13, no. 2, pp. 521–526, 2019.

[3] Jiawei Han and M. Kamber, Data Mining: Concepts and Techniques Second Edition. 2013.

[4] D. Patel, R. Modi, and K. Sarvakar, “A Comparative Study of Clustering Data Mining: Techniques and Research Challenges,” Int. J. Latest Technol. Eng. Manag. Appl. Sci., vol. 3, no. 9, pp. 67–70, 2014.

[5] P. Indirapriya and D. D. K. Ghosh, “A Survey on Different Clustering Algorithms in Data Mining Technique,” Int. J. Mod. Eng. Res., vol. 3, no. 1, pp. 267–274, 2013.

[6] J. De Weerdt, S. Vanden Broucke, J. Vanthienen, and B. Baesens, “Active trace clustering for improved process discovery,” IEEE Trans. Knowl. Data Eng., vol. 25, no. 12, pp. 2708–2720, 2013.

[7] M. Verma, M. Srivastava, N. Chack, A. K. Diswar, and N. Gupta, “A Comparative Study of Various Clustering Algorithms in Data Mining,” Int. J. Eng. Res. Appl. www.ijera.com, vol. 2, no. 3, pp. 1379–1384, 2012.

[8] V. W. Ajin and L. D. Kumar, “Big data and clustering algorithms,” in International Conference on Research Advances in Integrated Navigation Systems, RAINS 2016, 2016.

[9] A. Novikov, “PyClustering: Data Mining Library,” J. Open Source Softw., vol. 4, no. 36, p. 1230, 2019.

[10] D. Xu and Y. Tian, “A Comprehensive Survey of Clustering Algorithms,” Ann. Data Sci., vol. 2, no. 2, pp. 165–193, 2015.

[11] B. Karthikeyan, D. J. George, G. Manikandan, and T. Thomas, “A comparative study on k-means clustering and agglomerative hierarchical clustering,” Int. J. Emerg. Trends Eng. Res., vol. 8, no. 5, pp. 1600–1604, 2020.

[12] P. K. Jain and R. Pamula, “Two-Step Anomaly Detection Approach.”

[13] A. Saxena et al., “A review of clustering techniques and developments,” Neurocomputing, vol. 267, pp. 664–681, 2017.

[14] S. Rashid, A. Ahmed, I. Al Barazanchi, and Z. A. Jaaz, “Clustering algorithms subjected to K-mean and gaussian mixture model on multidimensional data set,” Period. Eng. Nat. Sci., vol. 7, no. 2, pp. 448–457, 2019.

[15] M. S. Michael Steinbach George, Vipin Kumar, “A Comparison of Document Clustering Techniques,” TextMining Work. KDD2000, pp. 75–78.

[16] M. V. Reddy, M. Vivekananda, and R. U. V. N. Satish, “Divisive Hierarchical Clustering with K-means and Agglomerative Divisive Hierarchical Clustering with K-means and Agglomerative Hierarchical Clustering,” Int. J. Comput. Sci. Trends Technol., vol. 5, no. Sep-Oct, pp. 5–11, 2017.

[17] H. H. Ali and L. E. Kadhum, “K- Means Clustering Algorithm Applications in Data Mining and Pattern Recognition,” Int. J. Sci. Res., vol. 6, no. 8, pp. 1577–1584, 2017.

[18] T. H. T. Nguyen, D. T. Dinh, S. Sriboonchitta, and V. N. Huynh, “A method for k-means-like clustering of categorical data,” J. Ambient Intell. Humaniz. Comput., no. Berkhin 2002, 2019.

[19] C. Yuan and H. Yang, “Research on K-Value Selection Method of K-Means Clustering Algorithm,” J, vol. 2, no. 2, pp. 226–235, 2019.

[20] T. Zhang and F. Ma, “Improved rough k-means clustering algorithm based on weighted distance measure with Gaussian function,” Int. J. Comput. Math., vol. 94, no. 4, pp. 663–675, 2017.

[21] O. M. Lior Rokach, “Clustering methods,” Adv. Inf. Knowl. Process., no. 9781447167341, pp. 131–167, 2015.

[22] S. D. Bin Dong, Zengyou He, Xiaofei Xu, “K-Histrograms: An Efficient Clustering Algorithm for Categorical Dataset*,” no. 1, pp. 6–8, 2003.

[23] A. Al Malki, M. M. Rizk, M. A. El-Shorbagy, and A. A. Mousa, “Hybrid Genetic Algorithm with K-Means for Clustering Problems,” Open J. Optim., vol. 05, no. 02, pp. 71–83, 2016.

[24] B. Zerhari, A. A. Lahcen, and S. Mouline, “Big Data Clustering : Algorithms and Challenges,” Proc. Int. Conf. Bihree Charact. Call. 3Vs (Volume, Veloc. Var. It Ref. to data that are too large, Dyn. complex. this Context. data are difficult to capture, store, Manag. Anal. using Tradit. data Manag., no. May, pp. 1–7, 2015.

[25] C. C. Aggarwal, Data classification: Algorithms and applications. 2014.

[26] M. Duan, K. Li, X. Liao, and K. Li, “A Parallel Multiclassification Algorithm for Big Data Using an Extreme Learning Machine,” IEEE Trans. Neural Networks Learn. Syst., vol. 29, no. 6, pp. 2337–2351, 2018.

[27] H. Azzag, G. Venturini, A. Oliver, and C. Guinot, “A hierarchical ant based clustering algorithm and its use in three real-world applications,” Eur. J. Oper. Res., vol. 179, no. 3, pp. 906–922, 2007.

[28] C. C. Aggarwal and P. S. Yu, “Outlier detection for high dimensional data,” SIGMOD Rec. (ACM Spec. Interes. Gr. Manag. Data), 2001.

A Genetic Algorithm Approach for Inter and Intra Homogeneous Grouping Considering Multi-student Characteristics

A. M. Aseere
King Khaled University
College of Computer Science
ABHA, The Kingdom of Saudi Arabia

***Abstract*—This paper addresses the problem of group formation in collaborative learning by considering the students' characteristics. The proposed solution is based on a Genetic Algorithm (GA), which minimizes an objective function that has two main aims. Indeed, the proposed GA's fitness function helps to achieve two objectives: Fairness in the formation of different groups, resulting in intergroup homogeneity, and a low gap in the levels of students within a group, which corresponds to intragroup homogeneity. Exhaustive experiments were conducted using three different sizes of randomly generated data sets and several crossover operators. Indeed, the order crossover and the crossovers based on random keys representation are experimented. The reported results show that the proposed approach guarantees the efficient grouping of students. In addition, comparisons with existing approaches based on GA confirm the ability of the proposed approach to provide greater intergroup and intragroup homogeneity. In addition, the uniform crossover based on random keys representation ensures better grouping quality than do the other experimented crossover operators.**

Keywords*—Genetic algorithm; group formation; intragroup homogeneity; intergroup homogeneity; fitness; permutation; random keys representation

I. INTRODUCTION

Group formation is a crucial issue in collaborative learning. At present, group work is becoming increasingly recommended in various disciplines. Group work is mandatory for the realization of projects that have a high workload, and which generally require different skills. It is well established that group work is extremely beneficial for the training of students, and allows them to improve the group-work skills that are highly sought by recruiters.

The task of forming groups is a very delicate one due to the disparity of students in terms of skills. The success of student groups relies on several factors, such as personality, expertise in performance, and the collaboration of the students within the group. In collaborative learning [1], experts recommend forming a group consisting of students with complementary skills while ensuring the narrowest possible gap between the levels of students. Very large differences among group members can hinder cooperation. In addition, when forming groups from a set of students, the grouping technique must ensure fairness between groups. In fact, the success of each student depends on the success of the group to which the student belongs.

Different approaches are presented in the literature to answer research questions related to the problem of group formation. The systemic literature review of the subject presented in [2] cites several works that answered the four important research questions related to the problem, namely the homogeneity/heterogeneity of groups, the learner characteristics considered, the ideal group size, and the techniques used for automated group formation. The works cited in [2] present different answers and points of view regarding each research question. The size of the groups may depend on the total number of students, the tutor's choice, or the supervisory capacity of the institution/faculty. In [3], it is mentioned that members become less productive when the group size is large. Thus, groups consisting of three to seven members may be reasonable. Whether the type of group is heterogeneous or homogeneous has been studied in several works. According to [2], most works have addressed heterogeneous grouping, and fewer were interested in homogeneous groups.

The heterogeneity of a group can assist students to achieve good results and can help them to learn how to collaborate and interact with different types of classmates. However, excessive differences among students in a group can hinder cooperation and collaborative learning [1]. In [4], authors present a literature review concerning the problem of group formation. Indeed, based on the studied literature, the authors provided taxonomies of the characteristics of group formation and surveyed the techniques for forming groups. Automatic heterogeneous or homogeneous grouping can be performed based on students' characteristics [5]. These characteristics could be academic (grades, tests, self-evaluations, and so on), cognitive (learning styles intelligence types, and so forth), personality traits (such as leadership skills), or other considerations. Various types of characteristics are used for automatic group formation based on different optimization techniques [2], [4]. In this paper, a specific objective function based on different types of characteristics is proposed for achieving the desired type of grouping.

GA based approaches have been used in several studies to accomplish automatic group formation [2], [4]. In [1], the authors proposed a two-step approach, namely a categorization step that aims to maintain intragroup diversity and intergroup balance, followed by an optimal formation step that uses a GA to provide an approximate solution for group formation after determining each student's category. In [6], an adaptive GA is employed to seek the best combinations of groups. The maximized fitness function is the sum of two terms: the balanced score based on score category assigned to the average scores of the students, and the inverse of the absolute distance between the average scores of groups. In [7], the maximally diverse grouping problem is addressed. The absolute distance between different attributes of various groups is maximized. A hybrid technique combining a local search algorithm and a GA was used to solve the posed problem. An approach based on GA was proposed to form heterogeneous groups [8]. The index of learning styles and the academic attributes of students were used. The proposed algorithm aims to minimize the distances among separate groups. Using GA approach to group formation considering multiple student characteristics was proposed [5]. In the aim to form inter and intra homogeneous groups, the considered computed fitness function of the GA corresponds to the squared differences with regard to all the characteristics considered for each group and the entire sample. However, the proposed fitness function focuses on intergroup homogeneity rather than on intragroup homogeneity. Since the average values of the characteristics of each group are considered, it may be difficult to avoid having a combination of students with high skills and others with very low ones within a group. In fact, the success of an unbalanced group is not guaranteed. In general, the students that are at higher levels take on most of the workload.

In this paper, a two-objective fitness function is proposed to achieve group formation with intragroup and intergroup homogeneity. The average levels of the groups must be very close, and the gaps in the levels of students within a group must be low. Thus, the proposed fitness function of the GA is the sum of two terms; the first term corresponds to the fitness function proposed in [5], and the second term maximizes the sum of the minimum of the characteristics across all the groups. As the algorithm computes the sum of the minimum characteristics for each group, and then attempts to maximize the minimum value of these sums, the aim is to improve the intragroup homogeneity.

The successful use of a GA relies on appropriate coding of the genotypes (chromosomes) and appropriate operators to generate a new population from the current one. Since each student can only be in one group, permutation can be any valuable solution for genotype coding. Nevertheless, specific crossover and mutation operators must be used for coding based on permutation. In [9], the authors showed that order crossover and interchanging mutation were best choices for permutation. However, in [10], random keys are presented as a robust representation technique that guarantees maintaining feasibility from parents to offspring. Thus, random keys representation makes it possible to use conventional crossover operators, such as a single-point crossover, a two-point crossover, and the like [11]. The random keys representation technique has been used successfully in job scheduling problem [12]. Permutation is used to code chromosomes, and the grouping corresponding to each permutation is done in a very similar way to the coding proposed in [13]. In [14] some recommendations based on exhaustive simulation results are provided for setting the values for the different parameters of a GA.

The remainder of the paper is organized as follows: Section 2 presents the mathematical formulation of the proposed group formation problem. The proposed fitness function is also defined. Section 3 describes the use of a GA, and presents the different operators used. Section 4 evaluates the performance of the proposed approach. The experimental protocol is presented, and the results obtained are reported and discussed. Finally, Section 5 provides some concluding remarks and recommendations for future work.

II. Problem Formulation

This paper proposes a method for combining a number of students into groups in the context of collaborative learning. The group formation is based on various characteristics of the students. The characteristics considered here correspond to the grades of the students in different modules such as Math, Programming Skills, Communications Skills, their GPAs, and so on. The characteristics considered for the groupings can be set by the administration of the faculty according to the requirements of the collaborative learning program.

This paper proposes a method for combining a number of students into groups in the context of collaborative learning. The group formation is based on various characteristics of the students. The characteristics considered here correspond to the grades of the students in different modules such as Math, Programming Skills, Communications Skills, their GPAs, and so on. The characteristics considered for the groupings can be set by the administration of the faculty according to the requirements of the collaborative learning program.

Let us assume that we have to group N students, and that $\mathcal{S} = \{s^1, s^2, \cdots, s^N\}$ is the set of students. Each student $(s^i)_{i=1,\cdots,N}$ is characterized by an array of L characteristics, denoted as $C^i = (c_1^i, c_2^i, \cdots, c_L^i)$. These characteristics could be of different forms, such as academic (grades, tests, self-evaluations, and so on), cognitive (learning styles, intelligence types, and so forth), personality (leadership, personality traits, and the like), or others. The numerical values of these characteristics may vary in different ranges. In such a case, the data must be normalized prior to the optimization process. Normalization aims to reduce the variations of all the characteristics' values to same interval.

Let us assume that $N_G = \frac{N}{P}$ is the number of groups, where P is the number of students per group. For each group $g = 1, \dots, N_G$, we associate a matrix M_g made of the arrays of the P students' characteristics. Thus, M_g is a $P \times L$ matrix defined as follows: $M_g = (C^{1\{g\}}, \cdots, C^{P\{g\}})^T$, where $(C^{i\{g\}})_{i=1,\cdots,P} \in \{C^1, C^2, \cdots, C^N\}$.

The idea is to distribute students within the groups in order to achieve two objectives:

1) Homogeneity between groups, and

2) A low gap in the levels of students within the same group.

With reference to [5], a criterion that focuses on achieving the first objective was proposed. In this paper, a modified objective function that helps to satisfy both objectives is proposed.

1) Intragroup homogeneity: The aim is to reduce the deviation between the average of each characteristic of a group and the average of each characteristic of all the students. As proposed in [5], the best grouping is one that minimizes the objective function f_1, defined as:

$$f_1 = \sum_{g=1}^{N_G}\left[(\overline{C_1} - \overline{M_{g,1}})^2 + (\overline{C_2} - \overline{M_{g,2}})^2 + \cdots + (\overline{C_L} - \overline{M_{g,L}})^2\right] \tag{1}$$

where $\overline{C_j} = \sum_{i=1}^{N} c_j^i$ and $\overline{M_{g,j}} = \frac{1}{P}\sum_{i=1}^{P} c_j^{i\{g\}}$ correspond respectively to the average of the characteristic j over N students, and the average of the characteristic j over P students in the group g.

2) Intragroup homogeneity: In order to guarantee intragroup homogeneity, maximizing the objective function f_2 is proposed as follows:

$$f_2 = \min_{g=1,\cdots,N_G}(\min(M_{g,1}) + \min(M_{g,2}) + \cdots + \min(M_{g,L})) \tag{2}$$

where $\min(M_{g,j}) = \min(c_j^{1\{g\}}, c_j^{2\{g\}}, \cdots, c_j^{P\{g\}})$ corresponds to the minimum of the characteristic j over P students in a group g. When attempting to minimize the function f_1, there is a significant chance of finding groups in which there is a large difference in the levels of students within the same group. This effect can be reduced by maximizing the function f_2 at the same time.

3) Fitness function of the GA: This paper uses the GA for the automatic grouping of students. For the posed problem, the GA seeks a set of feasible solutions for the optimal solution (the best grouping of the students) that minimizes a specific fitness function. The latter is defined in the following.

Considering equations (1) and (2), the fitness computed by the GA for each individual, i, in the population is defined as:

$$F^i = \beta\, f_1^i + \frac{1-\beta}{f_2^i + \varepsilon}, \tag{3}$$

where, f_1^i , f_2^i correspond respectively to the objective functions, f_1, f_2 are computed for the individual i, ε is a small parameter added to avoid division by 0, and $0 \le \beta \le 1$ is a weighting factor.

The best individual in a population is the one that has the lowest value of the fitness function, as defined in (3). The GA tends to decrease the lowest value of the fitness over a number of generations. In this paper, we propose attaching equal importance to f_1^i and f_2^i, so that β is fixed at 0.5. However, when β is set to 1.0, we retrieve the fitness function proposed in [5], defined as

$$F^i = f_1^i \tag{4}$$

III. Genetic Algorithm Deployment

For a given population (also called a generation), the classical GA executes the following steps in order:

1) The evaluation of the fitness of each individual in the population;

2) the selection of the best individual for breeding to create the new population; and

3) the application of the elitism principle followed by specific crossover and mutation operators to generate the new population.

In a conventional GA. the third step is constituted by the crossover step followed by the mutation step. However, in [15] an adaptation process is proposed, in order to adapt the individuals in the population to the best solution of the algorithm. In our work, we choose the elitism process to copy a few best solutions from the current generation to the future one.

Starting from an initial population, the algorithm repeats steps 1, 2, and 3 during several iterations. In each iteration, it creates a new population that may contain the best individual corresponding to the best grouping of students. The algorithm can be stopped by limiting the number of generations, by setting a threshold for the lowest fitness level, or by limiting the execution time of the algorithm. In general, the number of generations is used as the criterion for stopping. The coding of the individuals in the population, also called genotypes, depends on the posed problem, and is a key factor in selecting the specific genetic operators (crossover and mutation).

1) Genotype coding: Since each student can only be part of one group, it is a proposed that a permutation of the set of the students' indexes $\{1,2,\cdots,N\}$ be used as a genotype.

Let us consider a $\sigma = \{\sigma(0), \sigma(1), \cdots, \sigma(N-1)\}$, where $\sigma(i)$ is the integer at the i^{th} position of the permutation, and corresponds to the student $s^{\sigma(i)}$. Thus, each individual in the population is a permutation of integers from 0 to $N-1$, and the grouping of the students into groups of P students is performed as follows: The first group consists of the first P integers of the permutation $\sigma(0), \dots, \sigma(P-1)$, the second group is formed by $\sigma(P), \dots, \sigma(2P-1)$, and so on. In general, the i^{th} group is formed by $\sigma((i-1) \times P), \dots, \sigma(i \times P\ -1)$.

2) Crossover operators: The crossover operators used with permutations are extremely specific. It is not possible to use a conventional single-point crossover, a two-point crossover, and so forth [11] because it is difficult to maintain feasibility from parents to offspring. In [9], the authors showed that the order crossover was the best operator for permutations. In this paper, we experimented with other crossover operators based on the random keys representation presented in [10]. The idea was to use specific mapping to correspond to each permutation in a list of random numbers (keys). Thus, it was possible to apply a single-point crossover, a two-point crossover, or a uniform crossover to the lists of random numbers. After breeding, inverse mapping was applied to transform the lists of random numbers into permutations.

The use of order crossover and random keys representation for single-point, two-point, and uniform crossovers are presented in this paper.

3) Mutation operator: The interchanging mutation operator is recommended for genotypes coded as permutations [9], [5]. Thus, we propose to use the interchanging mutation with a low mutation rate [14].

4) Selection of individuals for breeding: The strategy based on elitism was proposed to create a new population from a current one, while maintaining the same size [11]. From the current population, α individuals with the best fitness were copied to the new population. The remaining individuals were obtained via the application of crossover operators on the selected parents for breeding. In fact, the selection of parents for breeding was performed via the roulette-wheel method based on the normalized cumulative of the fitness [11].

In this section, we investigate the performance of the proposed grouping approach based on the GA that minimizes the proposed fitness function, as defined in (3), for $\beta = 0.5$.

IV. Computational Experiments and Results Analysis

A. Experimental Protocol

Computational experience was carried out for three data sets that were generated randomly with three different sizes; small (N = 20), medium (N = 40), and large (N = 60). For all the data sets, the four characteristics that were considered (L = 4) were ‘Math’, ‘Programming Skills’, ‘Communication Skills’, and ‘GPA’. These characteristics may be required for groups working in a computer science collaborative learning program; for example, on a final graduation project. Since the values for the GPA are different from the values of the other characteristics, the normalization of the data set is mandatory prior to the optimization process.

The normalization approach proposed in [5] was adopted. Accordingly, each characteristic $\left(c_j^i\right)_{j=1,\cdots,L}$ of each student $(s^i)_{i=1,\cdots,N}$, was normalized according to the following procedure:

$$c_j^i = \frac{c_j^i - \left(c_j^i\right)_{\min}}{\left(c_j^i\right)_{\max} - \left(c_j^i\right)_{\min}} \quad (5)$$

where $\left(c_j^i\right)_{\min} = \min_{i=1,N}\left(c_j^i\right)$, and $\left(c_j^i\right)_{\max} = \max_{i=1,N}\left(c_j^i\right)$. Following this normalization, the values of the normalized characteristics varied in the interval of [0,1].

1) Setting the parameters of GA: Based on [14], for N students, the size of the population varied from N to 3N. In all the experiments, we set the size of the population as 2N. We conducted several experiments on the order crossover and the single-point crossover, the two-point crossover, and the uniform crossover based on random keys representation. The aim is to determine which one is the most appropriate for the proposed genotype coding. The mutation rate was fixed at 0.001 for all the performed experiments and the number of generations was the criterion for stopping the algorithm.

2) Assessment criteria: Obviously, the first criterion used to assess the performance of the proposed approach to group formation based on GA was the best fitness, F_{best}, which is computed as follows:

$$F_{\text{best}} = \min_{\text{generations}}\left(\min_{i\in\text{population}} F^i\right) \quad (6)$$

To evaluate the importance of the proposed fitness function (3) compared to the one defined by (4), we proposed computing the best grouping corresponding to F_{best} using the following parameters:

$$\text{Avg}_{\min} = \min_{g=1,\cdots N_g}\left(\sum_{j=1}^{L} \overline{M_{g,j}}\right) \quad (7)$$

$$\text{Avg}_{\max} = \max_{g=1,\cdots N_g}\left(\sum_{j=1}^{L} \overline{M_{g,j}}\right) \quad (8)$$

$$\text{Avg}_{\text{median}} = \text{median}_{g=1,\cdots N_g}\left(\sum_{j=1}^{L} \overline{M_{g,j}}\right) \quad (9)$$

The parameters (7), (8), and (9) characterize the degree of homogeneity between different groups, thus corresponding to the first objective. In fact, the higher the values of these parameters the greater the homogeneity between groups.

In addition to the parameters (7), (8), and (9) that evaluate the average values of the characteristics of the groups, and in order to evaluate the gaps in the levels of students within each group, we proposed computing for the best grouping corresponding to F_{best} using the following parameters:

$$\text{Min}_{\min} = \min_{g=1,\cdots N_g}\left(\sum_{j=1}^{L} \min(M_{g,j})\right) \quad (10)$$

$$\text{Min}_{\max} = \max_{g=1,\cdots N_g}\left(\sum_{j=1}^{L} \min(M_{g,j})\right) \quad (11)$$

$$\text{Min}_{\text{median}} = \text{median}_{g=1,\cdots N_g}\left(\sum_{j=1}^{L} \min(M_{g,j})\right) \quad (12)$$

Parameters (10), (11), and (12) correspond to the second objective. In fact, the higher these parameters’ values, the greater the intragroup homogeneity.

B. Results Analysis

In the first step in the performance analysis of the proposed approach for group formation, we conducted simulations to assess the importance of the fitness (3) ($\beta = 0.5$) compared to fitness (4) ($\beta = 1.0$). For these simulations, we considered three different sizes (20, 40, and 60) of the randomly generated data sets and the four different crossover operators. The number of generations was set at 200, the mutation rate at 0.001; the Elitism rate, α, was equal to 10% of the population size and finally, for the uniform crossover based random keys, the parameter of the binomial law was set at 0.7 [12].

The results obtained are presented in Table I, Table II, and Tab. III. for $N = 20$, $N = 40$, and $N = 60$, respectively. The reported results are the average over 20 runs. We fixed the population size at 2N [12] and the initial population is constituted by 2N randomly generated permutations of integer numbers from 0 to $\mathrm{N}-1$.

The results presented in Table I, Table II, and Table III show that the quality of the groupings obtained using the proposed approach was much more appropriate than was the one obtained when using the fitness function defined by (4) and proposed in [5]. This was valid for all the experimental crossover operators. In fact, the values of the parameters (6) - (12) assessing the quality of the groupings obtained via the proposed approach are higher than the values of those obtained via the approach based on the fitness (4). As expected, adding the maximization of the sum of the minimum characteristics of all the groups in the objective function produced a significant improvement in the quality of the groupings.

In addition, the obtained results show that the order crossover was not the most appropriate operator for population breeding for all the data sets in the proposed approach. In fact, the two-point crossover and the uniform crossover based on random keys representation appeared to be more efficient.

For a further investigation of the performance of the proposed approach, Table IV, Table V and Table VI show the simulation results obtained for the three data sets with 600 generations. For each data set, we considered the crossover operators that provided the best results after 200 generations (Table I, Table II, and Table III). The reported results were the average over 20 runs, and the GA used a population of size 2N for a data set of size N.

The results presented in Table IV, Table V and Table VI confirm the superiority of the proposed GA based on fitness (3) over the GA based on fitness (4). As expected, increasing the number of generations assisted the algorithm to achieve better results than those obtained with 200 generations for all the crossover operators considered. Based on these reported results, we can state that the proposed GA with the uniform crossover based on random keys representation ensures better grouping quality than do the other crossover operators.

TABLE I. Comparative Analysis of the Performance of Genetic Algorithm based on Fitness (3) versus Fitness (4) (Population Size = 40, Generations = 200, Mutation Rate = 0.001, α= 4)

Random Data Set of 20 Students								
Crossover	Fitness	$\mathrm{F}_{\mathrm{best}}$	$\mathrm{Avg}_{\mathrm{min}}$	$\mathrm{Avg}_{\mathrm{max}}$	$\mathrm{Avg}_{\mathrm{median}}$	$\mathrm{Min}_{\mathrm{min}}$	$\mathrm{Min}_{\mathrm{max}}$	$\mathrm{Min}_{\mathrm{median}}$
Order	(3)	0,37819	2,31503	2,58601	2,43335	1,57611	2,00153	1,69432
	(4)	0,05711	2,01043	2,5641	2,29812	1,00369	1,86938	1,49700
R. Keys Single-point	(3)	0,36566	2,33045	2,63612	2,46733	1,62994	2,07089	1,72847
	(4)	0,05715	2,06049	2,52151	2,28918	1,04332	1,84752	1,45798
R. Keys Two-point	(3)	0,35583	2,34723	2,62273	2,47493	1,63435	1,93480	1,75328
	(4)	0,04701	2,04859	2,50071	2,27728	1,03092	1,72159	1,43301
R. Keys Uniform	(3)	0,34555	2,2946	2,67124	2,51640	1,63116	2,06780	1,80386
	(4)	0,04712	2,07614	2,50701	2,30293	1,02322	1,81396	1,49647

TABLE II. Comparative Analysis of the Performance of Genetic Algorithm based on Fitness (3) versus fitness (4) (Population Size = 80, Generations = 200, Mutation Rate = 0.001, α = 8)

Random Data Set of 40 Students								
Crossover	Fitness	$\mathrm{F}_{\mathrm{best}}$	$\mathrm{Avg}_{\mathrm{min}}$	$\mathrm{Avg}_{\mathrm{max}}$	$\mathrm{Avg}_{\mathrm{median}}$	$\mathrm{Min}_{\mathrm{min}}$	$\mathrm{Min}_{\mathrm{max}}$	$\mathrm{Min}_{\mathrm{median}}$
Order	(3)	0,61242	2,04115	2,73540	2,27964	1,10697	2,19040	1,53068
	(4)	0,15982	1,91476	2,56528	2,20301	0,52619	2,01572	1,41403
R. Keys Single-point	(3)	0,58795	2,02214	2,61416	2,28030	1,16734	2,00016	1,48775
	(4)	0,19972	1,88904	2,53112	2,21207	0,64927	2,00527	1,36829
R. Keys Two-point	(3)	0,56642	2,01923	2,66400	2,29310	1,24155	2,10547	1,57232
	(4)	0,17039	1,90926	2,59815	2,26121	0,55352	2,04251	1,48417
R. Keys Uniform	(3)	0,58600	1,99183	2,64434	2,29607	1,14184	2,05865	1,49780
	(4)	0,17349	1,91332	2,54612	2,21842	0,62825	1,93114	1,37765

TABLE III. Comparative Analysis of the Performance of Genetic Algorithm based on Fitness (3) versus Fitness (4) (Population Size = 120, Generations = 200, Mutation Rate = 0.001, α = 12)

Random Data Set of 60 Students								
Crossover	Fitness	$\mathrm{F}_{\mathrm{best}}$	$\mathrm{Avg}_{\mathrm{min}}$	$\mathrm{Avg}_{\mathrm{max}}$	$\mathrm{Avg}_{\mathrm{median}}$	$\mathrm{Min}_{\mathrm{min}}$	$\mathrm{Min}_{\mathrm{max}}$	$\mathrm{Min}_{\mathrm{median}}$
Order	(3)	0,64552	1,92776	2,53707	2,23878	1,09650	2,07752	1,52831
	(4)	0,22394	1,82434	2,51451	2,18508	0,62140	2,02689	1,39662
R. Keys Single-point	(3)	0,68964	1,90218	2,58344	2,24333	1,05206	2,12256	1,52230
	(4)	0,30427	1,80189	2,52980	2,18551	0,59824	2,04232	1,41119
R. Keys Two-point	(3)	0,68300	1,88965	2,58525	2,22402	1,08020	2,12506	1,52543
	(4)	0,30175	1,84326	2,53305	2,17113	0,66259	2,02101	1,38816
R. Keys Uniform	(3)	0,66417	1,92619	2,56619	2,26264	1,15363	2,11301	1,56361

TABLE IV. Simulation Results for Two-Point Crossover and Uniform Crossover based on Random Keys Representation (Population Size = 40, Generations = 600, Mutation Rate = 0.001, α= 4).

Random Data Set of 20 Students								
Crossover	Fitness	F_{best}	Avg_{min}	Avg_{max}	Avg_{median}	Min_{min}	Min_{max}	Min_{median}
R. Keys Two-point	Fitness (3)	0,354453	2,319143	2,605539	2,462328	1,602172	2,014233	1,720088
	Fitness (4)	0,042657	2,080786	2,489469	2,283906	1,063818	1,756801	1,382646
R. Keys Uniform	Fitness (3)	0,34834	2,313022	2,648039	2,484857	1,598868	1,982582	1,766234
	Fitness (4)	0,034438	2,110269	2,548716	2,327067	1,043996	1,876398	1,49996

TABLE V. Simulation Results for Single-Point Crossover and Uniform Crossover based on Random Keys Representation (Population Size = 80, Generations = 600, Mutation Rate = 0.001, α= 8)

Random Data Set of 40 Students								
Crossover	Fitness	$\mathrm{F}_{\mathrm{best}}$	$\mathrm{Avg}_{\mathrm{min}}$	$\mathrm{Avg}_{\mathrm{max}}$	$\mathrm{Avg}_{\mathrm{median}}$	$\mathrm{Min}_{\mathrm{min}}$	$\mathrm{Min}_{\mathrm{max}}$	$\mathrm{Min}_{\mathrm{median}}$
Order	(3)	0,59161	2,02472	2,61923	2,30961	1,18193	2,00459	1,57977
	(4)	0,13544	1,92959	2,53868	2,24520	0,73276	1,96688	1,37998
R. Keys Two-point	(3)	0,57600	2,02383	2,61031	2,31733	1,21341	2,02746	1,53499
	(4)	0.17705	1.91723	2.50001	2.23312	0.69542	1.96258	1.40115
R. Keys Uniform	(3)	0,52780	2,06095	2,63358	2,31399	1,28536	2,04107	1,59964
	(4)	0.14670	1.93933	2.50318	2.24888	0.64332	1.99622	1.42142

TABLE VI. Simulation Results for Single-Point Crossover and Uniform Crossover based on Random Keys Representation (Population Size = 120, Generations = 600, Mutation Rate = 0.001, α= 12)

Random Data Set of 60 Students								
Crossover	Fitness	$\mathrm{F}_{\mathrm{best}}$	$\mathrm{Avg}_{\mathrm{min}}$	$\mathrm{Avg}_{\mathrm{max}}$	$\mathrm{Avg}_{\mathrm{median}}$	$\mathrm{Min}_{\mathrm{min}}$	$\mathrm{Min}_{\mathrm{max}}$	$\mathrm{Min}_{\mathrm{median}}$
R. Keys Two-point	Fitness (3)	0,63032	1,96704	2,591758	2,247427	1,13809	2,09044	1,590941
	Fitness (4)	0,19903	1,85369	2,529114	2,179855	0,623051	2,08470	1,353187
R. Keys Uniform	Fitness (3)	0,63269	1,95224	2,564983	2,270216	1,202023	2,13042	1,586091
	Fitness (4)	0,26463	1,78407	2,49969	2,172122	0,608483	2,03858	1,362632

V. Conclusion and Future Work

This paper proposes a GA approach considering multi-student characteristics for forming inter- and intra-homogeneous groups. The proposed fitness function aimed to achieve the two objectives of minimizing the difference between the average characteristics of each group and the average of the students' characteristics and maximizing the minimum of characteristics of all the groups. In fact, the second objective is of assistance in reducing the gaps in the levels of the students within the same group. Consequently, it improves the intragroup homogeneity, which the first objective does not achieve.

The reported simulation results show that the quality of the groupings improved remarkably when using the proposed fitness function compared to the quality obtained via the fitness function based only on the first objective. Since the genotypes were coded by permutation, the crossover operator based on the random keys representations was employed. Specifically, the uniform crossover provides better grouping quality than do the single-point crossover and the two-point crossover. In addition, the analysis of the obtained simulation results demonstrates the uniform crossover based on random keys representation is more efficient than is the order crossover that is generally used with permutations.

In the future, we plan to test the proposed approach on real data sets in order to further emphasize the performance of the proposed approach for group formation and, eventually propose further adjustments of the GA parameters.

References

[1] Dai-Yi Wang, Sunny S.J. Lin, Chuen-Tsai Sun, DIANA (2007). A computer-supported heterogeneous grouping system for teachers to conduct successful small learning groups. Computers in Human Behavior , 23, 1997–2010.

[2] Odo, C., Masthoff, J., & Beacham, N. (2019). Group formation for collaborative learning: A systematic literature review. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), Artificial Intelligence in Education: Proceedings of the 20th International Conference, AIED, Springer-Verlag, 2, pp. 206-212. https://doi.org/10.1007/978-3-030-23207-8_39.

[3] Kravitz, D. A., & Martin, B. (1986). Ringelmann rediscovered: The original article. Journal of Personality and Social Psychology, 50(5), 936–941.

[4] Maqtary, N., Mohsen, A. & Bechkoum, K. (2019). Group Formation Techniques in Computer-Supported Collaborative Learning: A Systematic Literature Review. Tech Know Learn, 24, 169–190.

[5] Julián Moreno, Demetrio A. Ovalle, Rosa M. Vicari. (2012) A genetic algorithm approach for group formation in collaborative learning considering multiple student characteristics. Computers & Education, 58(1), 560-569.

[6] P. I. Ciptayani, K. C. Dewi and I. W. B. Sentana. (2016). Student grouping using adaptive genetic algorithm. International Electronics Symposium (IES), Denpasar, 375-379.

[7] Fan, Zhi-Ping & Ma, Jian & Zeng, Shuo. (2011). A hybrid genetic algorithmic approach to the maximally diverse grouping problem. JORS, 62, 1423-1430.

[8] Anon Sukstrienwong, (2017). A Genetic-algorithm Approach for Balancing Learning Styles and Academic Attributes in Heterogeneous Grouping of Students. International journal of emerging technology in Learning, 12(3).

[9] Jihene Kaabi & Youssef Harrath. (2019). Permutation rules and genetic algorithm to solve the traveling salesman problem. Arab Journal of Basic and Applied Sciences, 26(1), 283-291.

[10] James C. Bean. (1994). Genetic Algorithms and Random Keys for Sequencing and Optimization. ORSA Journal on Computing, 6(2), 154-160. http://dx.doi.org/10.1287/ijoc.6.2.154.

[11] Reeves C. (2003). Genetic Algorithms. In: Glover F., Kochenberger G.A. (eds) Handbook of Metaheuristics. International Series in Operations Research & Management Science, 57. Springer, Boston, MA. https://doi.org/10.1007/0-306-48056-5_3.

[12] Lei, D. (2010). Solving fuzzy job shop scheduling problems using random key genetic algorithm. Int J Adv Manuf Technol, 49, 253–262.

[13] Chen, Rong-Chang, Chen, Shih-Ying, Fan, Jyun-You, Chen, Yen-Ting. (2012). Grouping Partners for Cooperative Learning Using Genetic Algorithm and Social Network Analysis. Procedia Engineering, 29, 3888-3893.

[14] Atidel Ben Hadj-Alouane, James C. Bean. (1997). A Genetic Algorithm for the Multiple-Choice Integer Program. Operations Research, 45(1), 92-101. http://dx.doi.org/10.1287/opre.45.1.92.

[15] R. Saraçoğlu and A. F. Kazankaya, "Developing an adaptation process for real-coded genetic algorithms," Computer Systems Science and Engineering, vol. 35, no.1, pp. 13–19, 2020.

The Effects of Privacy Preserving Data Publishing based on Overlapped Slicing on Feature Selection Stability and Accuracy

Mohana Chelvan P[1]
Department of Computer Science
Karpagam Academy of Higher Education (KAHE)
Coimbatore, India

Dr. Perumal K[2]
Department of Computer Applications
Madurai Kamaraj University
Madurai, India

***Abstract*—Feature selection is vital for data mining as each organization gathers a colossal measure of high dimensional microdata. Among significant standards of the algorithms for feature selection, the primary one which is currently considered as significant is feature selection stability along with accuracy. Privacy preserving data publishing methods with various delicate traits are analyzed to lessen the likelihood of adversaries to figure the touchy values. By and large, protecting the delicate values is typically accomplished by anonymizing data by utilizing generalization and suppression methods which may bring about information loss. Strategies other than generalization and suppression are investigated to diminish information loss. Privacy preserving data publishing with the overlapped slicing technique with various delicate ascribes tackles the issues in microdata with numerous touchy attributes. Feature selection stability is a vital criterion of data mining technique because of the accumulation of ever increasing dimensionality of microdata due to everyday activities on the World Wide Web. Feature selection stability is directly correlated with data utility. Feature selection stability is data centric and hence modifications of a dataset for privacy preservation affects feature selection stability along with data utility. As feature selection stability is data-driven, the impacts of privacy preserving data publishing based on overlapped slicing on feature selection stability and accuracy is investigated in this paper.**

***Keywords*—Overlapped slicing; privacy preserving data publishing; feature selection; Jaccard Index; selection stability**

I. Introduction

There will be a huge amount of high-dimensional microdata created by organizations because of regular exercises on online business, e-administration, and so on. The table that involves data with singular portrayal or single respondent but not aggregative data is called microdata [1]. In this data, each record has at least one delicate attribute and is independently elucidated in each record [2]. For most organizations, for getting an edge over the contenders, data mining which is the extraction of helpful information from the accumulated transactional data of officialdoms is typically used. Feature selection is a significant dimensionality decrease method in data mining that chooses the subset of pertinent traits. Precision, productivity, and model interpretability is improved by the application of the feature selection technique. Microdata publishing strategies including slicing, overlapped slicing, t-closeness, l-diversity, k-anonymity will bother the data to safeguard the privacy of data. This paper is connected with the effect on data utility along with feature selection stability in data mining by perturbation of dataset for the privacy preserving data publishing methods particularly slicing and overlapped slicing strategies.

II. Data Perspective Nature of Feature Selection Stability

Because of the progressions in Information Technology, the online assortment of microdata about people as high-dimensional datasets is the ordinary movement. In feature selection, just a subset of significant traits is acquired as a dimensionality decrease strategy to defeat the "curse of dimensionality". Regardless of whether there will be a little perturbation or expansion of new samples, there ought to be a selection of a comparative set of traits in feature selection. As a significant measure of algorithms for feature selection, feature selection stability is considered as the ensuing iteration of feature selection should choose a comparable set of traits. Else, it brings down their conviction of researchers, as it will make disarray in the scientist’s mind about the findings of their research discoveries. Prior research contributions are toward the path that feature selection stability is generally dependent on algorithms. In any case, late explorers have demonstrated that feature selection is dataset dependent yet not algorithmic free [3-8].

Salem Alelyani and Huan Liu suggest that changes to the characteristics of the data set can affect the feature selection stability [3]. Previous researchers recommended that the feature selection stability is usually algorithmic dependent. However, Salem Alelyani and Huan Liu have tried well that the feature selection stability is usually dataset-reliant, but not entirely algorithmic independent. Salem Alelyani, Zheng Zhao, and Huan Liu suggest that there is a difficulty in measuring the stability of feature selection algorithms [4]. In his doctoral thesis, Salem Alelyani investigated the causes of instability in high-dimensional datasets using well-known feature selection algorithms and proved that feature selection stability mostly depends on data [5].

In the well-known feature selection algorithms, the stability results of different high-dimensional datasets from different domains will not be the same [6]. Some algorithms work better than the other domain datasets for a particular domain. Noise sensitivity is a major concern for algorithm instability in the selection of features. In many fields, a large amount of high-dimensional data range from social media, e-commerce, bioinformatics, healthcare, and online education [7]. The underlying data characteristics have a significant impact on the stability of the feature selection algorithm, as the stability problem can also depend on the data and these factors include the dimensionality of the feature, the number of data instances, etc. While the supervised feature selection requires prior knowledge on the label of each data instance, a new sample that does not belong to any existing classes will be considered as an outlier and there is no need to change the selected feature set to adapt to the outliers [8]. In other words, unsupervised feature selection is more sensitive to noise and the noise affects the stability of these algorithms.

The discrepancy in the distribution of the dataset can influence the feature selection stability. Data variation will influence selection stability. Particularly for choppy datasets like privacy conserved datasets, feature selection stability is generally influenced as privacy safeguarding annoyance influences the qualities of the dataset as feature selection stability is data-driven.

III. Jaccard Index

There are different measures for feature selection stability. Spearman's and Pearson's correlations expect to gauge the consistency of ranks or weights of the two lists of traits while the Jaccard Index plans to assess the measure of crossover between two sets of trait indices. In the investigations, the Jaccard Index is utilized to gauge feature selection stability. The subsets of results that contain chosen traits' indices and the Jaccard Index assess the stability by assessing the measure of crossover between the subsets [9]. The Jaccard Index estimates the similitude between finite sample sets and is characterized as the size of the crossing point partitioned by the size of the association of the sample sets estimates. Jaccard Index for two chose subsets is shown by the accompanying equation (1). From the given various outcomes R = {R1, R2... Rl} relating to various folds of the data set D, its stability can be surveyed by the measure of crossover between the sets in R as in the equation (2).

$$S_j(R_{i,}R_j) = \frac{|R_i \cap R_j\ |}{|R_i \cap R_j\ |} \tag{1}$$

$$S_j(R) = \frac{2}{l\ (l-1)} \sum_{\mathrm{i}=1}^{l-1} \sum_{\mathrm{j}=\mathrm{i}+1}^{l} S_j(R_{i,}R_j) \tag{2}$$

when the Jaccard Index value is 0, which is the result in which the feature selection results are not steady and 1 is the result where the outcomes are indistinguishable, consequently truly stable. Hence the value of Jaccard Index SJ restores is in the interval of [0, 1].

IV. Privacy Preserving Data Publishing

There will be a huge accumulation of high dimensional private data in organizations as every day we carry out our work in online technologies because of advancements in web technologies. For settling on strategic decisions, data mining is basic for business organizations. Data mining has been generally done by people who are not working in the organizations thus safeguarding the privacy of data is significant. Microdata including data about people like clinical data are published for their utility in research works yet uncovering private data could influence the notoriety of the organization and will lead the hefty monetary misfortunes. Privacy preserving data publishing is to ensure the privacy of data by some path before publishing microdata to an outsider for data mining.

V. Privacy Preserving Approaches

A few privacy preserving approaches have been intended for microdata publishing, for example, swapping, suppression, perturbation, randomization, and sampling [10]. Tuples in a similar bucket can't be recognized by their quasi identifier traits by generalization which changes the quasi identifier ascribes in each bucket into "less explicit however semantically reliable" values. Suppression substitutes the recognizing traits with values like '*'. Generalization falls flat on high dimensional data because of the scourge of dimensionality and it causes a lot of information loss because of the uniform-dispersion suspicion. For anonymizing high-dimensional data, bucketization has been chiefly utilized. By arbitrarily permuting the delicate trait values in each bucket after parcels tuples in the table into buckets, bucketization isolates the quasi identifiers with the touchy trait.

VI. Privacy Threats

Attribute disclosure, membership disclosure, and identity disclosure are the dangers of microdata publishing in data mining. Anonymising data would bring about better insurance from these dangers. Identity disclosure is worried about the revelation of identifying traits of the distinct individual. For securing attribute disclosure, coordinating different buckets was significant [11]. If the selection criteria were not a delicate trait value, at that point, it would prompt have a membership disclosure as membership information would surmise an identity of a person through different assaults [12].

VII. Microdata Publishing Techniques

A. *k-anonymity*

There will be a chance of aberrant identification of records from public databases utilizing a quasi-identifier ascribes, and to tackle this k-anonymity model was created. In this strategy, utilizing generalization and suppression techniques, the granularity of data representation is lessened [13]. In the k-anonymity procedure, each mix of estimations of quasi identifier traits can be vaguely coordinated to in any event k respondents. The values of the traits are discretized into stretches for quantitative ascribes or assembled into various sets of values for categorical attributes [13]. Notwithstanding, this method is lacking to forestall attribute disclosure. The background knowledge attack and the homogeneity attack are the two assaults on k-anonymity.

B. *l-diversity*

The l-diversity model was intended to deal with certain shortcomings in the k-anonymity model. Particularly when there is a homogeneity of delicate values inside a gathering, the k-anonymity model doesn't ensure the touchy traits relating to the quasi identifier ascribes. To forestall the assaults on k-anonymity, for example, background knowledge attack and homogeneity attack, the idea of intra-bunch diversity of touchy values is advanced inside the anonymization plan with the bucketization method. There will be not just support of the base gathering size of k, yet besides centers around keeping up the diversity of the touchy traits, in the method of l-diversity [14]. If there are 'l well-represented' values for a delicate trait, the class is said to have l-diversity. It is the most straightforward comprehension of 'well-represented' when there will be a guarantee that there are in any event l distinct values for the touchy trait in every proportionality class [14].

C. t-closeness

All values of a given attribute along these lines independent of its dispersion in the data are in the case of the l-diversity model. In any case, for the condition for genuine data sets, the trait values might be quite slanted. Regularly, a foe may utilize background knowledge on the global appropriation to make derivations about delicate values in the data. The property utilized by the t-closeness model is that the separation between the global dispersion and the dissemination of the delicate trait inside an anonymized gathering ought not to be quite the same as by more than a threshold t and this is [15]. In comparing the numerous other privacy conserving data mining techniques for numeric ascribes, the t-closeness approach will in general be more successful. An equality class is said to have t-closeness if the separation between the spread of the trait in the entire table and the appropriation of a touchy trait in this class is close to a limit t.

D. Slicing

It doesn't matter for data, even that doesn't have a reasonable partition between quasi identifying traits and touchy traits, bucketization doesn't forestall membership disclosure. Particularly for high dimensional data, generalization loses an impressive measure of information. There will partition the data both on a level plane and vertically in the case of a slicing strategy. This method protects preferable data utility overgeneralization. Dividing ascribes into columns will secure privacy by transgressing the relationship of uncorrelated traits and safeguard data usefulness by saving the relationship amongst profoundly interrelated traits [11, 12]. Slicing is more practiced for high-dimensional data. The sliced table is shown in Table I.

E. Overlapped Slicing

An augmentation of slicing techniques is applied in overlapped slicing is. By putting the touchy trait into a quasi-identifier section after copying the touchy trait, the overlapped slicing is performed. The placing of a trait into more than one column is the thought behind overlapped slicing. In overlapped slicing, a touchy trait like sickness will be copied and placed into a quasi-identifier column [16]. Since there is more attribute relationship, this will build data utility. Consequently, Table II depicts overlapped slicing which is an expansion of Table I.

The disease has overlapped as in the first and second columns, which is shown in Table II. It will give better data utility by overlapping this trait. At that point, to keep up privacy ensure, the arbitrary permutation of tuples in each bucket from the second column is performed [17]. How arbitrary permutation is acted in overlapped slicing is portrayed below in Table III. It is not influenced by information loss as the counterfeit tuples are made by this arbitrary stage. Haphazardly permutation of values in each bucket in the second column takes place to break connecting between two columns, as shown in Table III.

Values in the primary bucket {((32, M, Teacher, Cancer)), ((30, M, Lawyer, Viral Infection)), (25, F, Clerk, Heart Disease), (31, F, Teacher, Cancer)} are haphazardly permutated as appeared in Table II, and the values {(25023, Viral Infection), (26364, Cancer), (26385, Heart Disease), (26895, Cancer)} are arbitrarily permutated. Additionally, in this way connecting between the two columns in a single bucket is camouflaged.

Quasi identifiers are not generalized or suppressed in overlapped slicing, as we find in Table III. The method of permuting arbitrarily delicate values in a bucket is shown in this technique. These will in general limit information loss if it requisite be generalized and hence this strategy doesn't create any information loss. Notwithstanding, counterfeit tuples are produced in this overlapped slicing. Touchy values permutation in a bucket results in the outcome. Since the information is complete, these phony tuples don't diminish the utility of data.

TABLE I. SLICED TABLE

(Zip Code, Disease)	**(Age, Sex, Occupation)**
(24281, Heart Disease)	(24, F, Police)
(26385, Heart Disease)	(25, F, Clerk)
(26895, Cancer)	(31, F, Teacher)
(26364, Cancer)	(39, M, Priest)
(23022, Cancer)	(32, M, Teacher)
(24102, Viral Infection)	(30, M, Lawyer)
(25023, Viral Infection)	(28, F, Teacher)

TABLE II. TABLE AFTER OVERLAPPED SLICING

(Zip Code, Disease)	**(Age, Sex, Occupation, Disease)**
(24281, Heart Disease)	(24, F, Police, Heart Disease)
(26385, Heart Disease)	(25, F, Clerk, Heart Disease)
(26895, Cancer)	(31, F, Teacher, Cancer)
(26364, Cancer)	(39, M, Priest, Cancer)
(23022, Cancer)	(32, M, Teacher, Cancer)
(24102, Viral Infection)	(30, M, Lawyer, Viral Infection)
(25023, Viral Infection)	(28, F, Teacher, Viral Infection)

TABLE III. OVERLAPPED SLICING TABLE WITH RANDOM PERMUTATION

(Zip Code, Disease)	**(Age, Sex, Occupation, Disease)**
(26895, Cancer)	(24, F, Police, Heart Disease)
(25023, Viral Infection)	(25, F, Clerk, Heart Disease)
(23022, Cancer)	(30, M, Lawyer, Viral Infection)
(24102, Viral Infection)	(32, M, Teacher, Cancer)
(26364, Cancer)	(28, F, Teacher, Viral Infection)
(26385, Heart Disease)	(39, M, Priest, Cancer)
(24281, Heart Disease)	(31, F, Teacher, Cancer)

VIII. EXPERIMENT

A. Methodology

The following gives the intended methodology for the privacy shielding algorithm i.e., slicing and overlapped slicing:

1) The ranking technique of information gain is utilized for picking quasi identifier traits.

2) Mean, variance and, standard deviation which are statistical properties for trial datasets are hounded.

3) CFS feature selection algorithm was utilized on trial datasets.

4) Before privacy ensuring ruffling, the accuracy for picked traits is hounded.

5) By the privacy conserving algorithm and hereby slicing and overlapped slicing, the quasi identifier ascribes, and touchy traits are disturbed.

6) Factual properties, for instance, mean, variance, and, standard deviation are hounded for privacy conserved datasets.

7) For datasets conserved for privacy, the CFS feature selection algorithm has been put on.

8) For the privacy conserved datasets, utilizing Jakkard Index JI, feature selection stability calculations are dogged.

9) After privacy preserving ruffling, the accuracy of picked attributes is hounded.

*10)*For the datasets conserved for privacy utilizing the privacy conservation algorithm i.e., slicing and overlapped slicing, accuracy, feature selection stability, and the statistical properties are dissected.

B. Information Gain IG

It is outlined as a method that provides additional Y data provided by X that displays the value by which the entropy of Y drops [18]. Entropy might be a basis for impurity in an actual training set S. This topic is assigned as IG, which is symmetrical and is determined in (3).

$$\text{IG} = \text{H(Y)} - \text{H(Y/X)} = \text{H(X)} - \text{H(X/Y)} \quad (3)$$

Knowledge picked up comparable to Y shadowed by the perception of X is equivalent to the knowledge picked up according to X, trailed by the perception of Y. This can be acclimated to help traits with high values, even though it will be not any more informative and is the deficiency of the IG rule. Taking into account the disparity amongst the entropy of the trait and along these lines, the conditional entropy gave the class label, the IG picks the opportunity inside a trait and the class label. It computes the value of a trait, considering the information gain upheld by the class as shown in (4).

$$\text{IG (Class, Attribute)} = \text{H (Class)} - \text{H (Class | Attribute)} \quad (4)$$

C. Correlation-based Feature Selection CFS

The preference is indicated for subsets of traits that are noticeably correlated inside the class with the inter-correlation among classes will be scanter [19]. By observing the level of redundancy among them, alongside the unmistakable prophetic capacity of each trait, the value of the attribute subsets is assessed by CFS. Authors have GA as a search strategy with CFS as a fitness function. CFS can be provided with elective search mechanisms and chooses the best trait subset. In (5) CFS is indicated.

$$r_{zc} = \frac{Kr_{zi}}{\sqrt{\text{K} + (\text{k}-1)r_{ii}}} \quad (5)$$

where, r_{zi} alludes to the mean connections between subset traits and the class variable, r_{ii} alludes to the mean inter-correlation between subset traits, and r_{zc} gives the relationship between different subsets of traits and the class variable, whereas k alludes to the number of subset traits [19].

D. Datasets Used

The two datasets utilized in the experimentations are the COIL 2000 - Insurance Company Benchmark dataset [20] and the KDD - Census-Income dataset [20]. The KEEL dataset repository is the source from which the datasets are accessible. Table IV shows the data sets highlights. Inside the chronicled data sets, the Coil 2000 dataset essentially has numerical values and the KDD dataset has each categorical and numerical value.

E. Experimental Results Analysis

By considering the status of traits, by utilizing the information gain concerning the class, the ranked traits are gotten. The Information Gain IG feature selection algorithm was utilized to accomplish this. The quasi identifier traits are gotten from which are maintained by the ranked ascribes and were chosen for slicing and overlapped slicing methods. For a hundred percent privacy conservation, every domain value of the picked trait has been perturbed. Even with critical background knowledge, an interloper or roguish data miner can't be sure about the exactness of a re-identification.

To pick ascribes from both original and privacy safeguarded data set, the CFS feature selection algorithm is applied. The CFS algorithm does not collaborate with any classifier inside the selection system as the algorithm is a filter-based. BestFirst is the search method utilized in the trial. By 10-fold cross-validation, overfitting is decreased. BestFirst is redesigned with backtracking abilities and it practices greedy hillclimbing to search for the region of trait subsets. BestFirst interest forward once it begins with the empty set of traits or requests toward each way anytime by requesting all single trait extensions and cancellations which are examined at a predetermined point or it can look for in switch once it begins with the total set of traits.

TABLE IV. FEATURES OF DATASETS COIL 2000 AND CENSUS

S. No.	Datasets Characteristics	Datasets	
		Coil 2000	**Census**
1	Type	Classification	Classification
2	Attribute Type	Numerical	Numerical, Categorical
3	Classes	2	3
4	Features	85	41
5	Origin	Real World	Real World
6	Mislaid Values	No	Yes
7	Instances	9782	143228

As the feature selection stability would increment be competent to up to the ideal number of relevant ascribes and after that decrease, the number of traits picked has been maintained in an idyllic assortment. For both the original and the bothered dataset numerical traits, the statistical properties, for instance, standard deviation, variance, and mean are assessed. In comparing with other privacy preserving data mining techniques, in the case of slicing and overlapped slicing methods, the statistical properties are not much affected. For the privacy safeguarded datasets, the undertakings for statistical exhibitions are engaged for validation.

For the dataset Coil 2000 and the dataset Census, feature selection stability is assessed with the stability measure Jaccard Index JI and the results have been revealed up in Fig. 1. As feature selection stability is data-centric, the discrepancy of the dataset, for instance, the privacy safeguarding perturbation is unfairly correlated with the feature selection stability along with the data utility.

The privacy preserving algorithms i.e., slicing and overlapped slicing has made stable feature selection results by considering the statistical properties of the numerical traits of the perturbed datasets which are stanch. The Census dataset has equally numerical and categorical traits, whereas the Coil 2000 dataset has all the numerical ascribes. Also, as it is found from the exploratory results, the Coil 2000 dataset is altogether more stable than the Census dataset, because it just contains numerical traits.

Because the feature selection stability results for the privacy safeguarding algorithms are effective, the accuracy of the privacy conserved datasets is nearly comparable to before modification. There will be a trade-off among accuracy, feature selection stability, and privacy safeguarding perturbation, and hence in the exploratory investigations, the accuracy results are given need alongside feature selection stability over different estimates, for example, ROC, F1-score, precision, AUC, runtime analysis. The data utility and feature selection stability are decidedly related. Thusly, data utility, feature selection stability, and privacy safeguarding have been tried for the slicing and overlapped slicing procedures with two unique datasets. At that point, the exploratory results showed that the activity of the data publishing techniques on investigational datasets prompts a stable feature selection alongside unassailable accuracy. Table V sums up the measures of the controlled experimentation on the test datasets close by the kinfolk which include feature selection stability along with accuracy.

The measures of the test results of the slicing and overlapped slicing are nearly the equivalent. By and by, the overlapped slicing method offers a marginally preferable exhibition over the slicing procedure, as appeared in the diagram beneath in Fig. 1.

TABLE V. FOR SLICING AND OVERLAPPED SLICING, ACCURACY AND FEATURE SELECTION STABILITY MENSURATION SUMMARY USING DATASETS COIL 2000 AND CENSUS

Investigational Outcomes	Slicing		Overlapped Slicing	
	Coil 2000	Census	Coil 2000	Census
Feature selection stability exploiting Jaccard Index JI	0.88	0.83	0.91	0.86
Overall accuracy before ruffling	74.93%	72.78%	74.93%	72.78%
Overall accuracy after ruffling	67.23%	66.12%	71.62%	68.73%
The Accuracy of chosen features before ruffling	80.79%	77.83%	80.79%	77.83%
The Accuracy of chosen features after ruffling	75.28%	74.12%	78.86%	76.72%

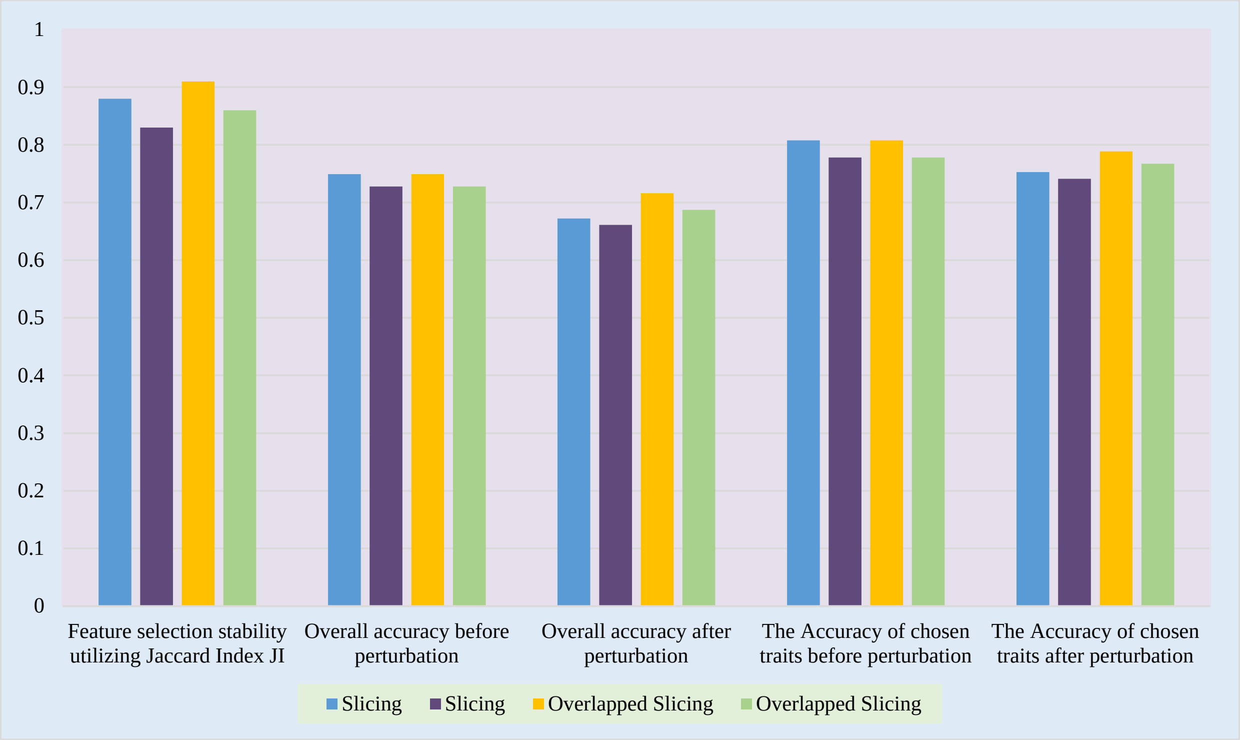


Fig. 1. For Slicing and Overlapped Slicing, Feature Selection Stability and Accuracy Mensuration using Datasets Coil 2000 and Census.

IX. CONCLUSION

This research contribution gives an overview of feature selection stability and its importance in data mining. It likewise examines different privacy preserving data publishing methods. It likewise gives a record of how the data publishing methods influence the selection stability and accuracy in privacy safeguarding data mining. Feature selection stability is now considered a significant criterion in data mining techniques especially due to the accumulation of private data with ever increasing dimensionality due to advancements in internet technologies and smartphones. The slicing and overlapped slicing are recent procedures of privacy preserving data publishing. The key findings of the experimental studies concluded the significance of the overlapped slicing technique in terms of feature selection stability along with data utility in comparison with other privacy preserving data publishing techniques. Overlapped slicing has preferred data utility over slicing strategy. The experimental analyses show that overlapped slicing is superior to the slicing technique regarding feature selection stability and accuracy.

REFERENCES

[1] Can, O. Personalised anonymity for microdata release. IET Inf. Secur. 2018, 2, 341–347.

[2] Taylor, L.; Zhou, X.H.; Rise, P. A tutorial in assessing disclosure risk in microdata. Statistics in Medicine; Wiley: Hoboken, NY, USA, 2018; pp. 1–14.

[3] Salem Alelyani, Huan Liu., The Effect of the Characteristics of the Dataset on the Selection Stability, IEEE DOI 10.1109/International Conference on Tools with Artificial Xiniun Intelligence, 2011.167, 1082-3409/11, http:// ieeexplore.ieee.org/document/ 6103458, 2011.

[4] Salem Alelyani, Zheng Zhao, Huan Liu., A Dilemma in Assessing Stability of Feature Selection Algorithms, IEEE DOI 10.1109/ International Conference on High Performance Computing and Communications, 2011.99, 978-0-7695-4538-7/11, http:// ieeexplore.ieee.org/document/6063062, 2011.

[5] Salem Alelyani, On feature selection stability: a data perspective, Doctoral Dissertation, Arizona State University, AZ, USA, ISBN: 978-1-303-02654-6, ACM Digital Library, 2013.

[6] Barbara Pes, Feature Selection for High-Dimensional Data: The Issue of Stability, Proceedings of the 26th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE 2017), June 21–23, 2017.

[7] Jundong Li, Huan Liu, Challenges of Feature Selection for Big Data Analytics, Special Issue on Big Data, IEEE Intelligent Systems, eprint arXiv:1611.01875, 2017.

[8] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu, Feature Selection: A Data Perspective. ACM Comput. Surv, 50, 6, Article 94, 45 pages. DOI: https://doi.org/10.1145/3136625, 2018.

[9] Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. Knowledge and Information Systems, 12(1):95–116, May 2007.

[10] Yan Zhao, Ming Du, Jiajin Le, Yongcheng Luo, "A Survey on Privacy Preserving Approaches in Data Publishing" in the First International Workshop on Database Technology and Applications, 2009.

[11] B.Vani, D.Jayanthi, "Efficient Approach for Privacy Preserving Microdata Publishing Using Slicing" IJRCTT, 2013.

[12] Tiancheng Li, Jian Zhang, Ian Molloy, "Slicing: A New Approach for Privacy Preserving Data Publishing" IEEE Transaction on KDD, 2013.

[13] Charu C. Aggarwal, ‘‘On k-Anonymity and the Curse of Dimensionality", Proceedings of the 31st VLDB Conference, Trondheim, Norway, pp.901-909, 2005.

[14] Ashwin Machanavajjhala, Daniel Kifer,Johannes Gehrke, Muthuramakrishnan Venkita Subramanian," ℓ-Diversity : Privacy Beyond K-Anonymity", Proc. International conference on Data Engineering.(ICDE),pp.24, 2006.

[15] Anil Prakash, Ravindar Mogili, ‘‘Privacy Preservation Measure using t-closeness with combined l-diversity and k-anonymity", International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARC SEE)Volume 1, Issue 8,pp:28-33, 2012.

[16] Widodo, Budiardjo EK, Wibowo WC. Privacy Preserving Data Publishing with Multiple Sensitive Attributes based on Overlapped Slicing. Information. 2019; 10(12):362.

[17] Widodo; Wibowo, W.C. A Distributional Model of Sensitive Values on p-Sensitive in Multiple Sensitive Attributes. In Proceedings of the International Conference on Informatics and Computational Science, UNDIP Semarang, Kota Semarang, Indonesia, 30–31 October 2018.

[18] Hall, M A., and Smith L A., Practical feature subset selection for machine learning, Proceedings of the 21st Australian Computer Science Conference, Springer.181- 191, 1998.

[19] Mark A. Hall, Correlation-based Feature Selection for Machine Learning, Dept of Computer science, University of Waikato, http://www.cs.waikato.ac.nz/ mhall / thesis.pdf, 1998.

[20] Alcalá-Fdez, A., Fernández, J., Luengo, J., Derrac, S., García, L. Sánchez, and Herrera, F., KEEL data-mining software tool: Dataset repository, integration of algorithms and experimental analysis framework, J. Multiple-Valued Logic Soft Comput. 17(2): 255–287, 2010.

Exploring UX Maturity in Software Development Environments in Saudi Arabia

Obead Alhadreti
Department of Computer Sciences, Computing College
Umm Al-Qura University, Al-Qunfudah, Saudi Arabia

***Abstract*—User experience (UX) design is becoming increasingly crucial for developing successful software today. It can determine whether or not users stay engaged with a product or service. It is, therefore, important that organizations have their users in mind when developing software and that there is a maturity for UX work. However, there are still organizations which do not value UX highly and where UX maturity is low. This paper reported the results of a survey of 75 practitioners working in software-development environments in Saudi Arabia. The survey was conducted in July 2020 and aimed to explore practitioners' perceptions of UX maturity, UX significance, and the challenges that face UX process in software development environments. The results show a higher than expected perception of organizational UX maturity amongst the practitioners surveyed, with the majority considering their organizations to be at an "Integrated phase". The degree of awareness of UX value was also found higher than anticipated. Furthermore, the study reveals important information about the most used UX methods as task analysis, prototyping, and heuristic evaluation. It also shows that UX assessment and user involvement being considered during different stages of product development, particularly in the prototyping phase. The major challenges that face UX process were found to be the need to improve UX consistency and the ability of teams and departments to collaborate.**

***Keywords*—User experience; UX maturity; Saudi Arabia**

I. Introduction

Organizations are increasingly recognizing the business value of hiring user experience (UX) professionals and incorporating UX design. Indeed, high-profile companies such as Apple and Google have incorporated UX design as a centerpiece of their success. The demand for UX specialists in the industry is leading to increasing numbers of related courses at universities around the world, and there are more and more people trained and capable of excellence in UX design. However, a belief in UX and having UX skills and resources available are not sufficient to guarantee meeting corporate goals for UX design. The UX of a product is not solely down to a UX designer; it is a result of how the organization as a whole executes on the product creation. There is, therefore, an implied sentiment that more mature UX leads to organizational success [1].

The information and communications technology (ICT) sector is growing fast in Saudi Arabia. It currently constitutes one of the Middle East’s largest markets for telecoms and information technology. The Saudi Arabian Communications and Information Technology Commission (CITC) has stated that spending in the ICT sector has reached approximately USD35 billion in 2020, a growth of approximately 8.3% over 2015, due to digital-transformation initiatives employed by a number of organizations around the country [2]. Digital services are also rapidly growing due to the increase in the Internet penetration rates and mobile phone usage. Approximately 29.9 million people in Saudi Arabia were using the Internet by the end of 2018. Mobile phone penetration is at 130%, with 42.5 million subscribers [3].

The Saudi Arabian Government’s National Transformation Program has also played a role in enabling the IT industry to enhance its contribution to the non-oil component of the country’s gross domestic product (GDP) [4]. There were 1650 IT companies in the country in 2003, encompassing home-grown businesses as well as multinationals and local subsidiaries, and this figure has increased substantially. Then, there was only a few local IT companies which involved themselves in system development, which might be due to individuals and organizations preferring outsourced solutions [5]. Today, there are indications of organizations having already moved away from complete reliance on outsourcing and becoming providers of a proportion of services and products. That could be down to Saudi Arabian businesses preferring IT services tailored to local requirements, which can be achieved by companies located inside the country. It has also been noted that organizations in both the public and private sectors have developed in-house technology development resources, which may explain why one of commonest jobs in IT currently is “software developer,” and why there is an expectation that the specialty will continue to have high demand [6]. Furthermore, there is an ever-increasing number of courses and educational programs in UX design on offer at Saudi universities and other educational institutions. Likewise, there are more and more research centers and research departments engaged with the UX field [7].

With these growth indicators of the software industry and an increasing UX attention in education establishments, there is surprisingly little data about where the UX field currently stands in Saudi Arabia, what impact the increasing attention on UX has had on IT development in practice, what challenges UX facing. It appears that no study inspected UX maturity in Arab countries, where software-development environments may have different cultural and organizational standards, and societies have different cultural and local requirements of software products. This paper presents and discusses results from a survey of practitioners' perceptions of

UX maturity, and UX significance, and the challenges that face UX process in software development environments in Saudi Arabia. The paper is structured as follows. Section II reviews existing literature, with attention given to recent studies around UX maturity models. Later sections present the study's methodology, data analysis and its results. The final section sets out the conclusions drawn from the study.

II. Related Work

Studies of UX maturity have mainly focused on Europe and the USA. Rukonić, de Meerendré, and Kieffer [1] assessed four dimensions of UX maturity and capability, namely methods, resources, artefacts, culture and literacy. As a demonstration of how their model could be used, the researchers undertook a case study across four corporations. The data collection was carried out by means of a survey, which was made up of questionnaire, completed on-line, and interviews conducted remotely. A selection of the survey participants was interviewed to corroborate their survey answers, establish their level of UX literacy, and elicit their opinions of the value of UX in assisting with product development. There was a particular focus on how return on the investment in UX was understood by the participants and on how UX discipline was viewed by their employers. The results seem to accurately capture the current UX capability of an organization.

Anchahua, Garnique, and Tarazona [8] developed a UX maturity model that provides tools, practices and techniques to enhance user satisfaction and corporate revenue. To test the model, the authors used a checklist of 25 tools, applied to four web-based software applications, to assess their value in enhancing the user experience. Half of these tools demonstrated greater maturity, as was also seen in the measures of user satisfaction and usability. The model was subsequently tested on an e-commerce website, which initially scored maturity of 38%. This score rose to 67.5% following the application of the checklist.

Another study, Möller [9], studied an organization which found UX difficult to accommodate within its work routines. This was in part due to using agile methods of software development at the same time. Nielsen's eight-stage model was used to test the organization's UX maturity. A survey with questions based on the model was used to collect data. A series of five interviews, run in a semi-structured format, was used to follow this up. The study results indicated that the organization had reached no higher than stage four maturity, and that was in projects where there was a plan and budget for UX and where teams had specific UX roles. Other teams had a lower maturity level. The authors then discussed ways in which the organization could improve its UX maturity. The actions recommended consisted of planning for UX, finding a way to show UX results, allowing UX experts to take the lead, involving actual users, and structuring the work on UX around agile processes.

The first empirical study of UX maturity across diverse people and organizations was that carried out by Sauro, Johnson, and Meenan [10]. The authors described the first steps of a maturity model based upon empirical evidence and what they learned from surveys of practitioners in many organizations and from a series of interviews with experienced UX professionals. This empirically based methodology was shown by the results to be sufficiently flexible and intelligible for application in a broad spectrum of sectors and organizations. The assessment tool itself was a survey employing eleven measures:

- characteristics of the individual user;
- characteristics of the organization;
- staff involved in UX;
- methods for UX research;
- corporate culture and leadership;
- degree of integration of UX;
- skills and training related to UX;
- product success;
- business success;
- the budget and resources for UX; and
- challenges associated with and future directions for UX.

Young, Chao, and Chandler [11] perform a mixed-methods study of maturity of UX practice in academic libraries. Both qualitative analysis of the content and statistical analysis were applied to data from a survey carried out amongst UX practitioners. The results showed the extent and types of UX methods used by practitioners at the present time in academic libraries. Responses to the survey also allowed the extraction of a series of themes that reveals the factors influencing UX maturity. The authors' analysis looked in particular at the corporate characteristics influencing methods and maturity of UX. A maturity scale tailored for libraries was thereby developed and recommendations were made towards practices that could enhance academic libraries' UX maturity.

It is worth mentioning that, although there is variation in the names and structures of models designed to evaluate UX maturity, these models generally comprise between five and seven levels, which go from "unrecognized" up to "institutionalized," following a pattern along the lines of the pioneering "Organizational Human-Centeredness Scale" developed by Jonathan Earthy [12]:

1) Unrecognized
2) Recognized
3) Considered
4) Implemented
5) Integrated
6) Institutionalized.

Prior studies have also investigated perceptions of the significance of user-centered design (UCD) and UX amongst practitioners. Vredenburg, Mao, Smith, and Carey [13] surveyed UCD practitioners and reported that a large proportion of them supported the statement that "UCD ways of working had produced significant impacts in product development, as well as enhancing their companies' products' usability and usefulness". Ji and Yun [14] agreed with [13] that the UCD and usability practitioners surveyed supported the perception of UCD methods becoming more popular and more likely to be widely adopted in the future. Likewise, in a study of the Norwegian software development industry, respondents agreed that the success of their companies' products depended upon usability [15]. Similar conclusions were drawn from another study of current Malaysian practices in UX and usability. However, whilst many respondents agreed on the importance of UX and usability, they held usability to be more important than UX [16]. There has only been one study in Saudi Arabia of how IT professionals perceive HCI, that by Majrashi and Al-Wabil [7]. The results of that study showed a higher than expected recognition of the importance of HCI. Despite a growing body of literature on UX design and UCD methodologies in research contexts in the Arab world, the understanding of the practitioners' perspective remains limited.

Only a limited amount has been written about the challenges faced by UX processes in countries other than the USA and in Europe. Indian UCD practice was discussed by Henry [17]. According to him, the filed in the country is facing the same usability misconceptions that exist in some other parts of the world. However, he highlighted three main myths that he saw as responsible for the most damage to software development in India:

- "Pretty screens are all you need;"
- "I can design on my own; just give me some guidelines;" and
- "Usability is about testing."

IT development in Korea has also encountered similar misconceptions, according to Ji and Yun [14], which led to resistance to adopting rigorous user research or considering UCD/usability studies in the design process. As far as Saudi Arabia is concerned, it is not clear what challenges face UX practice. A key objective of the present study was to identify these UX challenges.

III. Method

A. Survey

The study was conducted using the questionnaire as a data gathering tool. The choice of a questionnaire as the most appropriate data collection tool for this study was made because it allows large volumes of data to be collected quickly, and it is widely recognized by respondents and can be administered easily [18]. The questionnaire used in this study was designed based on that used by Sauro, Johnson, and Meenan [10], mentioned earlier, and comprised 30 questions in eleven sections. The first section focused on the respondents' demographics and experience with UX (i.e., age, gender, educational background, current role with their employer, and number of years' experience with UX). The second section included questions about the company the respondents working at (e.g., staff numbers and product type). The remaining nine sections corresponded to the other nine domains set out by Sauro, Johnson, and Meenan [10]. The format of the questions varied, including, for instance, multiple choice and Likert-type scales. It was important to ensure the order of presentation of questions did not affect responses (e.g., survey fatigue can result in lower ratings in later questions). Accordingly, the order in which sections three to eleven was presented was randomized.

B. Sampling

One of the challenges most often reported in this type of research is the identification of organizations engaged in software development. The study population was defined as all organizations with an involvement in the development of software (public or private, professional or not) in Saudi Arabia. The respondents targeted were those IT practitioners in such organizations, whose jobs involved software-development environments.

There has been no previous attempt to establish a population of organizations involved with software development, so that respondents were contacted via various channels. The invitation and questionnaire were also posted on social media to reach further IT practitioners, particularly those in private companies with no mailing list accessible. The invitation emphasized that it was aimed at IT practitioners working for an organization involved in software development. It also defined UX and provided a link to more information on the concept. The invitation also advised potential respondents that they would be involved in the study voluntarily and that they would be able at any time to withdraw their participation. To encourage responses, respondents was not asked to identify the name of their organization, as respondents may have some concern about releasing specific information about their organization.

Both questionnaire and invitation were circulated in English as well as Arabic, to reach a wider population and achieve a higher response rate. Approximately half of the ICT professionals in the sector in Saudi Arabia are not nationals of the country [7]. In the event, the English version was used by approximately 34% of those who responded. To guarantee clarity, both versions (English and Arabic) were tested by three people as a pilot, with no issues identified. Google Forms was used to host the questionnaire on-line.

IV. Results and Discussion

A. Individual Profiles

75 respondents provided complete questionnaires. To identify if there was more than one response from the same organization, respondents' e-mail addresses was checked for those who used their organization's e-mail servers and respondents' answers were compared to specific questions in the organization profile section. Overall, it was confirmed that respondents could be from at least 73 different organizations. It was also determined that some of the large organizations, especially in the public sector, were represented in the sample based on the organization e-mail address provided voluntarily by some respondents.

66.66% of respondents were between 30 and 39 years of age. The remainder (33.34%) were between 21 and 29. Approximately 11% were female. All of the respondents were qualified to Bachelor's degree level, with major subjects given as computer science (50 respondents), behavioral science (17 respondents), and design (8 respondents). The job roles reported were product or project manager (26 respondents), developer (16 respondents), information architect (10 respondents), UX researcher (8 respondents), UX/UI designer (8 respondents), and other (7 respondents). There was a fair experience of UX reported (Fig. 1). The results demonstrate a good level of UX understanding among practitioners, which means that it would be worth considering their assessments of their organizations' level of UX maturity.

B. Organizational Profiles

The organizations represented in the sample varied in size, from just a few employees to more than 255 (see Table I). Only five respondents were employed by specialist UX companies, which suggests that there is a limited number of firms in the country which specialize in UX or consult in the field.

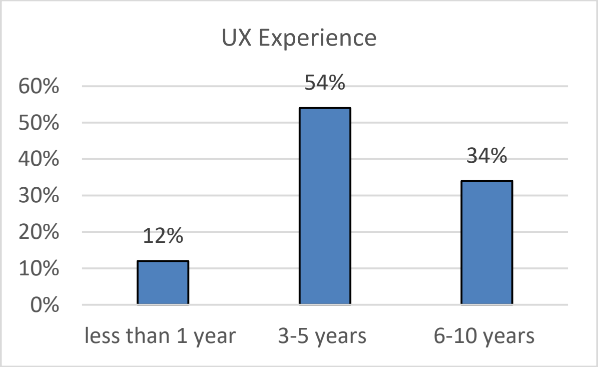


Fig. 1. Respondents' UX Experience.

TABLE I. ORGANIZATION PROFILES

Sector	Government	29
	Private	36
	Semi-government	10
Category	Software	26
	Education	9
	Health and medicine	7
	E-commerce	9
	Usability and UX	5
	Retail	4
	Other	15
Size	Large: 250 + employees	33
	Medium: 50–249	25
	Small: 10–49	11
	Micro: 1–9	6
Type of products	B2B	28
	Internal tools	14
	Consumer products or services	63
Platform	Mobile application	63
	Web or cloud-based	49
	Desktop software	42
	Hardware or physical products	7

The projects run by the organizations represented in the sample were software products or services aimed at consumers. This corroborates the shift in Saudi Arabia, reported by CITC, to in-house IT development departments, from reliance on outsourced provision and the adaptation of software available commercially [5]. A variety of platforms are employed in the organizations represented, but mobile, web and desktop applications predominate. The diversity seen in organizational profiles seems to correspond to Saudi Arabia's wide organizational range.

C. Estimated Maturity

56% of respondents in the sample perceived organization to be at the "Integrated" level of UX maturity, where UX processes are consistently integrated within product development (Fig. 2). It is interesting that no respondents considered their organizations to be at the "Unrecognized" level, where UX is not considered an issue and the user interface is mostly designed by developers, nor the "Ad hoc" level, where UX as an issue is recognized, but UX methods are not consistent.

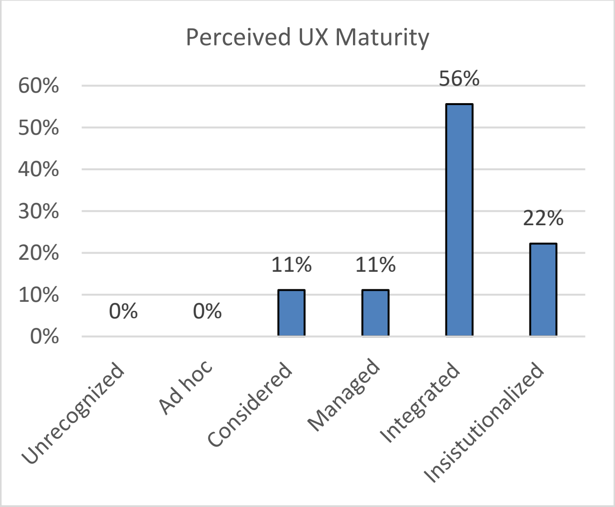


Fig. 2. Respondents' Perceptions of Organisational UX Maturity.

D. UX Staffing

Developers were better represented than either designers or researchers across organizations (see Fig. 3). 33% of respondents indicated a ratio of approximately 1 designer to 6-10 developers, and 22% reporting 1 to every 11-20. There was even more consensus when comparing designers to researchers, with 33.3% of respondents indicating a ratio of 1 researcher to every 5 or fewer designers. This finding is in line with that of [10], who assess UX maturity in US and found that there are approximately 1 designer to 11-20 developers, 1 researcher to every 5 or fewer designers.

E. UX Research Methods

Respondents were presented with a list of thirty UX methods and activities and asked to indicate which of these were employed in the software development environments within their organizations. Definitions of the thirty methods were provided for respondents. These were mostly taken from the literature (such as Hudson [19]). The respondents were also free to report methods additional to those given on the questionnaire. The methods and the frequency of their use was reported are shown in Table II. Although IT professionals seem to have some familiarity with various UX methods, the most frequently reported methods used by respondents were prototyping, task analysis, and heuristic evaluation.

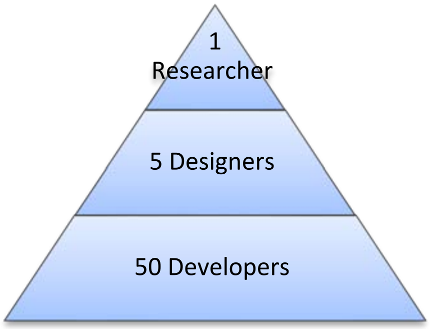


Fig. 3. Most Frequent Ratio of Researchers to Developers to Designers.

TABLE II. UX METHODS USED IN SOFTWARE-DEVELOPMENT ENVIRONMENTS

No.	UX methods	Frequency
1	Accessibility expert reviews	12
2	Accessibility testing	8
3	Analyze web metrics/logs	14
4	Benchmarking or competitive studies	24
5	Call center user feedback	22
6	Card sorting	25
7	Conceptual design	8
8	Content strategy	18
9	Content creation	16
10	Contextual inquiry	9
11	Creating prototypes (high-fidelity)	52
12	Creating prototypes (wireframes or low-fidelity)	45
13	Ethnography	0
14	Eye tracking	11
15	Focus groups	16
16	Heuristic or expert review	26
17	Information architecture	18
18	Interface / interaction design	33
19	Market research	15
20	Personas & user profiles	28
21	Project management	17
22	Requirements gathering	28
23	Satisfaction surveys	8
24	Strategy or strategic consulting	3
25	Surveys or other online research	11
26	Task analysis	41
27	Technical writing	1
28	Usability testing	16
29	User research (e.g. interviews & surveys)	14
30	UX / Design Workshops	12

Table III has been adapted from the work of [7] and [14] in order to illustrate the ten UX methods most used, according to the present study and other similar studies. Of these top ten methods used in Saudi Arabia, three have also been identified elsewhere in studies of IT-development practitioners, notably Korea [14], Europe and the USA [13], and across 14 countries (including the USA) [19]. The ways in which UX methods were applied to products was interesting to observe. Even though most respondents (61%) reported that they had a standards set of methods, these were applied flexibly according to the needs of the project in hand (see Fig. 4).

Respondents were asked for the UX resources available for use in their organizations. The most frequently reported resources were user panel access and the provision of prototypes for participants (see Table IV).

TABLE III. THE 10 MOST USED UX METHODS

Rank	Our study results	Ji and Yun [14]	Vredenburg et al. [13]	Hudson [19]
	Saudi Arabia	**Korea**	**US and Europe**	**US and other countries**
1	Prototyping	Task analysis	Iterative design	Informal usability testing
2	Task analysis	Evaluation of existing system	Usability evaluation	User analysis/profiling
3	Interface / interaction design	User analysis/profiling	Task analysis	Evaluation of existing system
4	Personas & user profiles	Surveys	Informal expert review	Low-fidelity prototyping
5	Requirements gathering	Heuristic/Expert evaluation	Field studies (contextual inquiry)	Expert (heuristic) usability evaluation
6	Heuristic or expert review	Scenarios of use	Focus groups	Task identification
7	Benchmarking or competitive studies	Navigation design	Formal heuristic evaluation	Navigation design
8	Card sorting	Usability checklists	Prototype without user testing	Scenarios of use
9	Call center user feedback	Focus-group interviews	User interviews	Set usability requirements
10	Information architecture	Lab usability testing	Surveys	Visual interface design

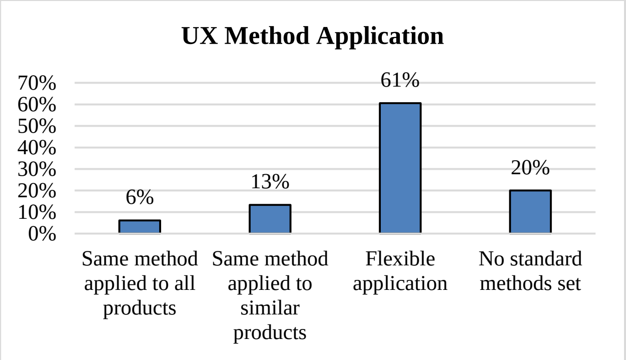


Fig. 4. UX Method Application.

TABLE IV. UX RESOURCES

UX resources	Frequency
A dedicated space for usability testing (e.g., research lab)	18
A one-way mirror for observation	3
Video feeds for observation	20
Access to user participant panels	58
Prototypes to provide to participants	56
Communication tools (intercom, instant messaging, etc.)	33

F. Leadership and Culture

Participants' perceptions of their organizations' valuation of UX were elicited by asking them to rank the perceived value on a scale from one to seven, where the higher the number, the higher the perceived value. The results appear to indicate that UX's importance is recognized by IT practitioners: there was an above-average score from all respondents. 22% gave the highest score, 33% at level six. In the present study, it was observed that respondents who reported their estimation of UX maturity higher, also reported a higher perceived value of UX, demonstrating that buy-in throughout the organization is strongly linked to maturity.

G. UX Integration/Application

One of the key indicators of an organization's level of UX maturity is the level of integration UX has within the organization. The present study's results reveal that UX is most frequently assessed at the stages of prototyping and advanced design, and after product launch (see Table V). The degree of UX maturity influences UX practices in an organization as well. Those respondents who gave higher maturity ratings for their organizations also reported UX evaluation at more stages.

Respondents were also asked how end-users were involved in product development. The most frequent stage for end-user involvement was prototyping (see Table VI). Significantly, the involvement of end-users at the prototyping stage was at a higher level in organizations rated higher for maturity.

H. Training and Skills

Respondents were also asked concerning the time and funding provided for UX education and training by their organizations. Almost 67% of participants reported having no funding available, and 78% reported having no time provided by their employer for UX education or training.

I. Business and Product Success Metrics

Respondents were asked to indicate the level of success experienced by their organization on a scale of one to seven, with higher numbers for more successful organizations. It is of interest that 75% of respondents perceived their organizations to be neutral in terms of success, 25% above average. It is possible that this is attributed to the time the survey was carried out, during the COVID-19 pandemic, which has affected businesses and other organizations very badly. Respondents were also asked to rate product development success, on a scale of one to five. Half believed that more than 50% of products made it at least as far as launch (see Fig. 5).

J. UX Budget and Resources

About 56% of participants reported having a dedicated budget for UX, whereas 33% reported having none. A few reported not having a dedicated budget, but that discretionary funding could be requested.

K. Challenges and Future Directions

The questionnaire's final section asked about challenges currently facing UX. Table VII shows the obstacles and how often they were reported. The issues raised most frequently were those of UX consistency enhancement and collaboration across departments and between teams. It is, therefore, vital to follow specific, established rules of design in order guarantee a product has a seamless experience overall. Furthermore, it is important for IT managers to ensure that UX's importance and its impact are clear to their teams. Respondents from more than 75% of organizations represented in the sample reported a lack of UX training to an adequate level. It is clear that more UX training programs need to be provided for practitioners. It would also be appropriate to provide developers with training tailored to the context of their language and culture, so that they can understand and become proficient in UX and UCD. Notwithstanding this, respondents did express optimism for UX's future, feeling that UX budgets would increase and that job opportunities in the field would expand.

TABLE V. UX ASSESSMENT DURING PRODUCT DEVELOIPEMNT

Stages of product	Frequency
Concept generation	32
Initial design/development	26
Prototyping/advanced design	68
Immediately prior to launch	12
Post-launch	66

TABLE VI. USER INVOLVEMENT DURING PRODUCT DEVELOIPEMNT

Stages of product	Frequency
Concept generation	29
Initial design/development	35
Prototyping/advanced design	40
Immediately prior to launch	9
Post-launch	7
Never	4

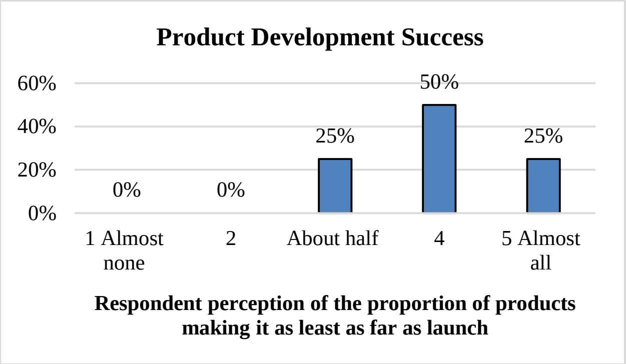


Fig. 5. Success of Product Development.

TABLE VII. CHALLENGES FACING UX PROCESS

UX challenges	Frequency
Handing off designs to developers	13
Collaborating between teams	34
Clarifying requirements	6
Securing appropriate UX budget or resources	17
Testing designs with end-users	20
Improving UX consistency	55
Legacy technology	19
Getting buy-in or understanding from executives	4
Collaborating across departments	42
Evolving UX and design alongside products (multiple testing time points)	4

V. CONCLUSION

Successful software development is becoming increasingly dependent upon UX design. It affects employee productivity, and whether a service or a product retains user engagement. Organizations must, therefore, ensure the maturity of their UX work. Yet, there are still organizations which do not assign high value to UX and where there is limited UX maturity. This paper reports the results of a survey of the current state of UX maturity in Saudi Arabian software development environments, involving 75 practitioners. A range of aspects was covered. In general, a higher awareness of the value of UX than expected was found. Practitioners also expressed a higher perception of their organizations' UX maturity than anticipated, with most considering it to be at an 'integrated' level. The study also reveals the UX methods most frequently employed (i.e., task analysis, prototyping and heuristic evaluation). The results also show that UX evaluation and the involvement of users occur at various product development stages, especially at the phase of prototyping. The main challenges for UX process were identified as collaboration across departments and between teams, and improving the consistency of UX. Future work is still required to confirm the results of the current study and reveal any other obstacles facing UX work in software development environments in Saudi Arabia.

ACKNOWLEDGMENT

The author would like to thank all those who take part in the survey. Thanks also the anonymous reviews for their invaluable comments.

REFERENCES

[1] Rukonić, Luka, Vincent Kervyn de Meerendré, and Suzanne Kieffer. "Measuring UX capability and maturity in organizations." In International Conference on Human-Computer Interaction, pp. 346-365. Springer, Cham, 2019.

[2] Communications and Information Technology Commission (CITC): annual report of Communications and Information Technology Commission (2016). 2016. Avaialbe at: https://www.citc.gov.sa/en/mediacenter/annualreport/Documents/PR_REP_012Eng.pdf. Accessed 22 July 2020.

[3] CITC: Indicators ICT KSA: end of Q3. 2018. Avaialbe at: https://www.citc.gov.sa/en/reportsandstudies/indicators/Indicators%20of%20Communications%20and%20Information%20Techn/ICTIndicators-Q32018En.pdf Accessed 22 July 2020.

[4] Saudi Vision 2030: National transformation program 2020. Avaialbe at: https://vision2030.gov.sa/sites/default/files/attachments/NTP%20English%20Public%20Document_2810.pdf Accessed 22 July 2020.

[5] CITC: IT report (2009). 2009 Avaialbe at: https://www.citc.gov.sa/en/mediacenter/annualreport/Documents/PR_REP_001E.pdf . Accessed 22 July 2020.

[6] CITC: ICT workforce report: 2015 (2015). http://www.citc.gov.sa/en/reportsandstudies/Reports/Documents/ICTWorkforce_en.pdf . Accessed 22 July 2020.

[7] Majrashi, Khalid, and Areej Al-Wabil. "HCI Practices in Software-Development Environments in Saudi Arabia." In International Conference on Cross-Cultural Design, pp. 58-77. Springer, Cham, 2018.

[8] Anchahua, Maritza Cieza, Luis Vives Garnique, and Javier Alvarez Tarazona. "User Experience Maturity Model for Ecommerce Websites." In 2018 Congreso Internacional de Innovación y Tendencias en Ingeniería (CONIITI), pp. 1-6. IEEE, 2018.

[9] Möller, Josefine. "Actions for Increasing an Organization's UX Maturity." (2018).

[10] Sauro, Jeff, Kristin Johnson, and Chelsea Meenan. "From snake-oil to science: measuring UX maturity." In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, pp. 1084-1091. 2017.

[11] Young, Scott WH, Zoe Chao, and Adam Chandler. "User Experience Methods and Maturity in Academic Libraries." Information Technology and Libraries 39, no. 1 (2020).

[12] Earthy, Jonathan. "Usability maturity model: Human centredness scale." INUSE Project deliverable D 5 (1998): 1-34.

[13] Vredenburg, Karel, Ji-Ye Mao, Paul W. Smith, and Tom Carey. "A survey of user-centered design practice." In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 471-478. 2002.

[14] Ji, Yong Gu, and Myung Hwan Yun. "Enhancing the minority discipline in the IT industry: A survey of usability and user-centered design practice."International Journal of Human-Computer Interaction 20, no. 2 (2006): 117-134.

[15] Bygstad, B., Ghinea, G., Brevik, E.: Software development methods and usability: perspectives from a survey in the software industry in Norway. Interact. Comput. 20(3), 375–385 (2008). https://doiorg.sdl.idm.oclc.org/10.1016/j.intcom.2007.12.001.

[16] Hussein, Idyawati, Murni Mahmud, and Abu Osman Md Tap. "A survey of user experience practice: a point of meet between academic and industry." In 2014 3rd International Conference on User Science and Engineering (i-USEr), pp. 62-67. IEEE, 2014.

[17] Henry, P.: Advancing UCD while facing challenges working from offshore. Interactions **10**(2), 38–47 2003. https://doi-org.sdl.idm.oclc.org/10.1145/637848.637861.

[18] Lazar, Jonathan, Jinjuan Heidi Feng, and Harry Hochheiser. Research methods in human-computer interaction. Morgan Kaufmann, 2017.

[19] Hudson, W.: Toward unified models in user-centered and object-oriented design. In: Van Harmelen, M. (ed.) Object Modeling and User Interface Design: Designing Interactive Systems, pp. 313–362. Addison-Wesley Longman, Boston. 2001.

Noise and Restoration of UAV Remote Sensing Images

Asmala Ahmad[1], Khadijah Amira Mohd Fauzey[2]
Mohd Mawardy Abdullah[3]
Centre for Advanced Computing Technology (C-ACT)
Fakulti Teknologi Maklumat Dan Komunikasi (FTMK)
Universiti Teknikal Malaysia Melaka, Melaka, Malaysia

Mohd Yazid Abu Sari[5]
Anjung Technology Sdn. Bhd
Ayer Keroh,
Melaka, Malaysia

Suliadi Firdaus Sufahani[4]
Faculty of Applied Sciences and Technology
Universiti Tun Hussein Onn Malaysia, Johor, Malaysia

Abd Rahman Mat Amin[6]
Faculty of Applied Sciences
Universiti Teknologi MARA
Terengganu, Malaysia

***Abstract*—Remotely sensed images captured from a camera mounted on a UAV (unmanned aerial vehicle) are exposed to noise caused by internal factors, such as the UAV system itself or external factors such as atmospheric conditions. Such images need to be restored before they can undergo further processing stages. This study aims to analyse the effects of salt and pepper noise on a UAV image and restore the image by removing the noise effects. In doing so, a UAV image, with red, green and blue channel and containing regions of different spectral properties, is experimented with salt and pepper noise of different densities. Image restoration procedure is formulated using median filtering of variable sizes. Peak-signal to noise ratio (PSNR) and mean square error (MSE) analysis are performed to measure image quality before and after restoration. An optimal filter size is chosen based on the highest PSNR of the restored image. The results show that the effects of noise on UAV images are dependent on the spectral properties of the image channels and the regions of interest. The proposed restoration works best for images with low- compared to high-density noises. Blue channel is found having the largest variation of optimal filter size, 18.5, compared to other channels because of the high response to noise within its short spectral wavelength region. Landscape's vegetation has the largest variation of optimal filter size, 22, compared to other regions due to the sensitivity of its dark spectral properties.**

***Keywords*—Noise; restoration; remote sensing; UAV; PSNR**

I. Introduction

Remote sensing technology has long been used for various purposes due to its spatial, spectral and temporal capabilities [1]. Initially, in the 1980s, satellite remote sensing technology was used to monitor various land covers continuously and offers cheaper costs compared to traditional approaches. Among the frequently used remote sensing satellites include Landsat, SPOT, IKONOS and Quickbird. These satellites have been actively used worldwide in numerous applications for more than 30 years; nevertheless, satellite imagery suffers limitations in terms of spatial and temporal resolution. Moreover, the satellite systems are operated by satellite operators in developed countries in which users do not have any autonomy over them. Other than that, images are sometimes unavailable for certain places and time besides exposed to other crucial issues, particularly cloud and haze effects [2]. Later, in the 1990s, besides satellite-based remote sensing technology, there were efforts to mount imaging systems on aircraft due to the need to capture images with higher spatial and temporal resolution as such onboard NASA Dryden DC-8 aircraft during the AirSAR PacRIM campaign; however, the operational costs were very expensive due to involving aircraft system maintenance [3]. Beginning 2000s, Unmanned Air Vehicles (UAVs) have been used to overcome the issues of using satellite- and aircraft-based technology; however, UAVs were initially massive and expensive in which the owners were normally among big organisations. Later in 2010s UAVs were then becoming affordable to many as well as smaller and lighter. Nowadays, a standard UAV is already equipped with an RGB camera and can be navigated autonomously. It has been currently used in various sectors related town planning, security, hazard monitoring, agriculture, environmental management and many more [4], [5]. Nevertheless, UAV-based remote sensing images tend to be exposed to noise from various internal or external factors [6]. Internal factors are caused by the UAV system itself including electronic and mechanical aspects of the UAV, while external factors are due to environmental issues such as haze, rain and fog. Such noise tends to modify the spectral properties and eventually degrades the UAV images qualitatively and quantitatively [7]. There exist studies on noise removal from satellite images however, comparatively, UAV images have significantly distinct spatial and temporal characteristics due to the different altitude and revisit frequency [7], [11], [17]. Therefore, this study attempts to investigate the effects of salt and pepper noise on a UAV image in which eventually, a noise removal procedure is to be proposed.

II. Evolution of Remote Sensing

Remote sensing satellites has initially been used to monitor various land covers due to its capability to capture images of large-scale agricultural land continuously and at an affordable cost. Among the frequently used remote sensing satellites include Landsat, SPOT, IKONOS and Quickbird. These satellites have been used actively worldwide for more than 30 years. Remote sensing satellites were designed with multispectral sensors to enable efficient monitoring of various Earth's resources however, satellite imagery suffers limitations in terms of spatial and temporal resolution. Spatial resolution can be defined as the ability to separate details in an image. Technically, spatial resolution is a measure of the smallest object that can be resolved by the sensor, or the ground area covered by the instantaneous field of view (IFOV) of the sensor [8]. Temporal resolution is a measure of the repeat cycle or frequency with which a sensor revisits the same part of the Earth's surface. The frequency characteristics are determined by the design of the satellite sensor and its orbit pattern. The spectral, spatial and temporal resolution of different remote sensing satellites are given in Table I.

In land cover monitoring, remote sensing satellites have sufficient spectral resolution to efficiently providing images of large areas, the drawbacks are in terms of spatial and temporal resolution. Certain objects such as building structures and road signs and crops are small wherein the size is far less than the IFOV of satellite sensors and therefore detection and monitoring could not be performed. As an example, crops such as paddy has very small leaves which contains important information to indicate the conditions of the plant however this is undetectable using satellites images [9]. Example GeoEye satellite images for paddy area in Kedah, located in the north-west of Peninsular Malaysia snipped from Google Maps are shown in Fig. 1. The left image is at a scale of approximately 1: 10,000 while the zoom-in image of the same spot is on right. Paddy leaves are still undetectable although after zooming-in the image to the highest level (right) [10].

In terms of temporal resolution, the frequency of satellite image acquisition depends very much on the satellite orbit and altitude. It ranges from once every 3 days to once every 16 days. The date and time of satellite overpass are fixed in which these satellites are operated by satellite operators in developed countries wherein users do not have any autonomy over them. The images are sometimes unavailable for certain places and time besides exposed to other crucial issues especially cloud and haze effects [7]. Here we display visibility data from Petaling Jaya located in Selangor, Malaysia, to demonstrate haze occurrence in Malaysia [11]. Fig. 2 shows a plot of daily visibility against day from 1999 to 2008. White, yellow, green, violet and red colours indicate clear (above 10 km visibility), moderate (5 – 10 km visibility), hazy (2 – 5 km visibility), very hazy (0.5 – 2 km visibility) and extremely hazy (less than 0.5 km visibility) conditions respectively (Table II). For most years, a drop in visibility can be observed at the end of the year, indicating the occurrence of increased haze.

Beginning 2010s, UAVs have become affordable to many and the size becomes smaller. Nowadays, a standard UAV such as DJI Mavic Pro has the dimension of 83mm x 83mm x 198mm (height x width x length), weighted only 743 g and is already equipped with RGB camera. Such UAVs are being currently used in various applications related to surveying and planning, agriculture management, security and many more [12]. In term of spatial resolution, certain objects, or targets have dimension mm to cm in size. To meet this requirement, the spatial resolution of a standard UAV need can be varied by changing the flying altitude. However, in term of temporal resolution, a standard UAV has limited endurance or flying duration per battery and also the ability for the propeller motors to withstand the produced heat due to the frequent and robust flying behaviour. Also, a standard UAV can fly only in 20 to 30 minutes per battery. Other than that, a standard UAV has a fixed imaging system that cannot be changed or customised. Due to such situations, we have developed an improved version of UAV known as Personal Remote Sensing System (PRSS) aiming to address the issues of standard UAVs. With PRSS, image capturing can be done automatically based on the pre-set flying waypoints. Besides that, PRSS battery and motor usage also have been improved where flying can reach 40 to 50 minutes. The mounted imaging system also can be changed and customised accordingly based on users' needs.

TABLE I. SPECTRAL, SPATIAL AND TEMPORAL RESOLUTION OF DIFFERENT REMOTE SENSING SATELLITES

Remote Sensing Satellite	Spectral Resolution (No. of bands)	Spatial Resolution (meter)	Temporal Resolution (days)
Landsat	6 (0.45 – 2.35 µm)	30	16
SPOT	4 (0.45 – 0.89)	10	5
IKONOS	4 (0.45 – 0.85)	3	3
Quickbird	4 (0.45 – 0.9)	2.4	3
GeoEye-1	4 (0.45 – 0.92)	1.6	8

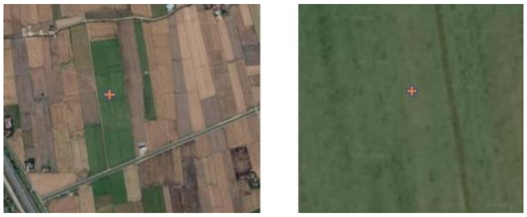


Fig. 1. Land Cover Classification using ML, NN and SVM when using 10% and 90% Training Set Sizes.

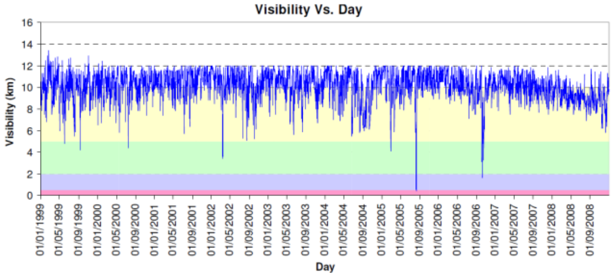


Fig. 2. Visibility against Day for Petaling Jaya from 1999 to 2008.

TABLE II. SPECTRAL, SPATIAL AND TEMPORAL RESOLUTION OF DIFFERENT REMOTE SENSING SATELLITES

Severity	Horizontal Visibility (km)
Clear	> 10
Moderate	5 – 10
Hazy	2 – 5
Very hazy	0.5 – 2
Extremely hazy	< 0.5

PRSS also aims to tackle the delay and cost issues related to traditional remote sensing data, improve the spatial and spectral resolution of the traditional remote sensing system and provide a user-friendly remote sensing system that can be used by anyone, at any place and in any time (Fig. 3). This system consists of a quadrotor UAV, a laptop as a processing unit and a smartphone for controlling and tracking. The quadrotor UAV is mounted with an RGB camera or any other imaging system to suit user's need. The UAV is equipped with GPS and telemetry facilities for tracking and controlling purposes. The camera is chosen to have GPS capabilities to provide a geographical location to the captured images. The acquired images are stored in the laptop and ready to undergo subsequent processing and analysis tasks in various applications. Fig. 4 shows the conceptual implementation of the PRSS.

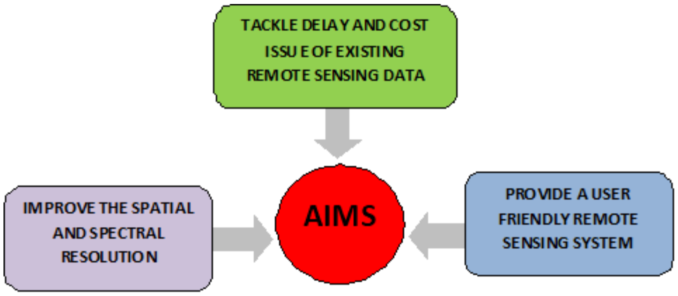

TACKLE DELAY AND COST ISSUE OF EXISTING REMOTE SENSING DATA
IMPROVE THE SPATIAL AND SPECTRAL RESOLUTION
AIMS
PROVIDE A USER FRIENDLY REMOTE SENSING SYSTEM

Fig. 3. Aims of PRSS.

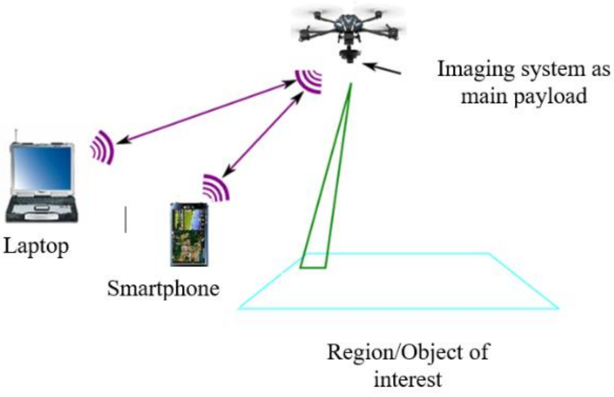


Fig. 4. Implementation of PRSS.

III. Noise and Restoration

We can consider a noisy image to be modelled as follows [13]:

$$g(x,y) = h(x,y) * f(x,y) + \eta(x,y) \qquad (1)$$

where $f(x,y)$ is the original image pixel, $h(x,y)$ is the degradation function, $\eta(x,y)$ is the noise term and $g(x,y)$ is the resulting noisy pixel. The objective of restoration is to obtain an estimate of the original image, $f(x,y)$. Following this, the model of image degradation and restoration can be illustrated in Fig. 5 [13]. The model indicates that if the model of the noise in an image can be estimated, then it is possible to figure out how the restoration of the image can be carried out.

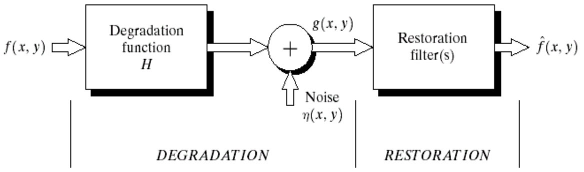


Fig. 5. Model of Image Degradation and Restoration.

In this study, impulse noise is to be chosen for the purpose of image degradation. There are three types of impulse noise which are salt noise, pepper noise and salt and pepper Noise. In an 8-bit image, salt noise is added to an image by addition of random bright (with 255 pixel value) all over the image. On the other hand, pepper noise is added to an image by addition of random dark (with 0 pixel value) all over the image. While the combination of both, salt and pepper noise is added to an image by addition of both random bright (with 255 pixel value) and random dark (with 0 pixel value) all over the image [15]. In other words, salt and pepper noise is a type of impulse noise and indicated by random black and white dots that appear in an image. This type of noise appears in the image due to the sharp and sudden changes of pixel’s brightness in an image. Consequently, an image that is affected by the salt and pepper noise contain obvious dark pixels in bright regions while bright pixels in dark regions [14]. Fig. 6 shows a test pattern image and the same image after salt and pepper noise is added together with their corresponding histogram [13]. In this study, the salt and pepper noise is to be simulated on a UAV image and a where its effects are to be assessed via PSNR and MSE.

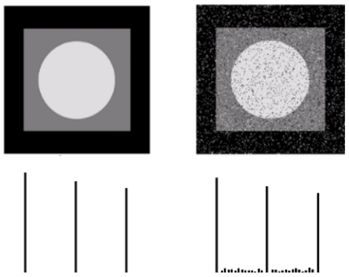


Fig. 6. A Test Pattern Image (Top Left) and the Same Image after Salt and Pepper Noise is Added (Top Right) and their Corresponding Histogram.

For the purpose of restoration, median filtering is to be used due to the ability to replace the grey level of each pixel by the median of the grey levels in a neighbourhood of the pixels [16]. Median filtering is an order statistics filtering process where it produces a restored image that is given by:

$$\hat{f}(x,y) = median\{g(x,y)\} \qquad (2)$$

where, $\hat{f}(x,y)$, the filtered image depends on the ordering of the pixel values of the noisy image $g(x,y)$, in the filter window. For a higher density of salt and pepper noise, the neighbourhood that form the window of a median filter can be enlarged by increasing the size of the median filter to effectively remove the noise [17]. In real remote sensing applications, such noise tends to cause errors in subsequent tasks such as land cover classification, object and feature detection and land surveying [18], [19], [20], [21], [22], [23]. To measure the quality of a corrupted image with respect to the noise-free image, peak signal-to-noise ratio (PSNR) can be used, in which PSNR is an engineering term for the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. Due to the different dynamic ranges of images, PSNR is frequently expressed in terms of the logarithmic decibel scale. The PSNR of a noisy image can be expressed as:

$$PSNR_g = 10log_{10}\left(\frac{MAX_f^2}{MSE_g}\right) \qquad (3)$$

$$MSE_g = \frac{1}{mn}\sum_0^{m-1}\sum_0^{n-1}[f(x,y) - g(x,y)]^2 \qquad (4)$$

where, MSE_g is the mean squared error of the noisy image, $f(x,y)$ is the noise-free image, $g(x,y)$ is the noisy image, m and n are the number of rows and columns of the image respectively. In the same way, the PSNR of a restored image can be expressed as:

$$PSNR_{\hat{f}} = 10log_{10}\left(\frac{MAX_f^2}{MSE_{\hat{f}}}\right) \qquad (5)$$

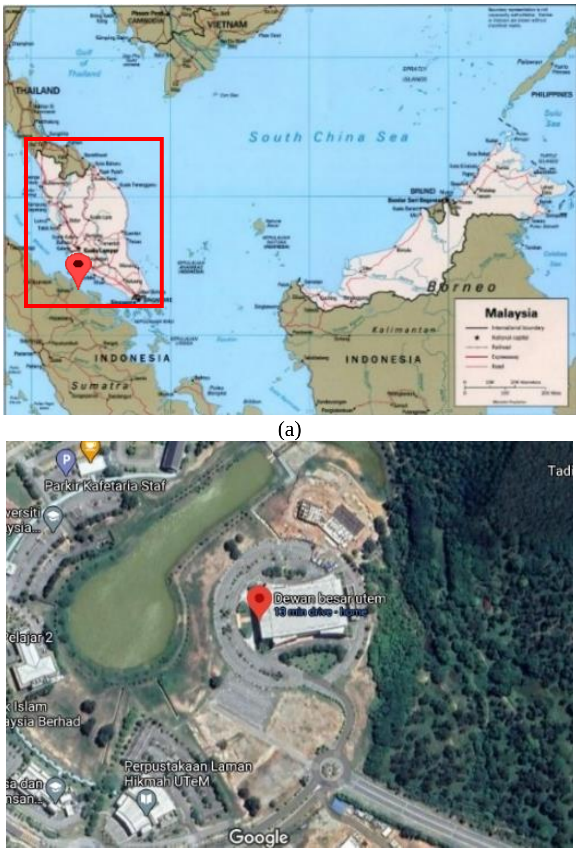
$$MSE_{\hat{f}} = \frac{1}{mn}\sum_0^{m-1}\sum_0^{n-1}\left[f(x,y) - \hat{f}(x,y)\right]^2 \qquad (6)$$

where, $MSE_{\hat{f}}$ is the mean squared error of the restored image, $f(x,y)$ is the noise-free image, $\hat{f}(x,y)$ is the restored image. These concepts will be adopted in understanding the effects of salt and pepper noise in a UAV image and formulating its restoration procedure.

IV. Materials and Methods

In this study, the experiment site is the main campus of Universiti Teknikal Malaysia Melaka (UTeM) located in Melaka, Malaysia. The main data come from PRSS imagery while ancillary data come from Google Maps. Fig. 7 shows the location of the study site, the grand hall of Universiti Teknikal Malaysia Melaka (UTeM), observed from the map of Malaysia, and its close-up from Google Maps. The main imagery used in this study was captured using a Canon PowerShot S100 camera that is mounted on the PRSS in which the implementation of image acquisition has been illustrated in Fig. 4. The image acquisition date was 27 March 2016 and the time was 9.37 am. It was a sunny day, and the sky is clear. Mission Planner software was used to plan the waypoints for PRSS navigation. Before taking off, the PRSS needed to be calibrated by turning the PRSS based on the x, y and z-axis which representing "roll", "pitch" and "yaw" accordingly. This is to ensure the PRSS getting a good set of roll, pitch and yaw tuning parameters for stable and accurate flight navigation. Next, the PRSS need to be “armed” and “disarmed” several times to ensure remote controlling was working properly. PRSS was then launched and the navigation was set to be “automatic” so that the PRSS follows autonomously the pre-set waypoints. The flying and battery conditions were tracked and monitored closely using the Mission Planner software. At the same time, the capturing of images was automatically performed based on the pre-set timing of the camera allowing automatic image acquisition to be performed along the waypoints for 40 to 50 minutes. Upon completing the mission, the PRSS was landed using the “return to home” function. Soon after safely landed, the PRSS was “disarmed” to power off the PRSS and the four rotors and battery were taken off. The acquired images that were saved in the camera’s storage card were downloaded into the laptop and are ready to be processed and analysed. MATLAB software was used for the purpose of image processing and analysis. Initially, images that contain important landmarks were sorted and an image which contains the grand hall of UTeM was chosen due to having the criteria needed for this study. Fig. 8 shows the UTeM’s grand hall captured from PRSS while the attribute and the corresponding metadata of the image are given in Table III.

From this image, four regions sizing 100 rows by 100 columns were subsetted: (a) grand hall’s roof surface, (b) landscape’s vegetation, (c) road surface and (d) balcony’s roof surface to represent the bright, dark, moderate dark and moderate bright condition respectively. This was to enable the effects of noise on different image condition to be investigated. The investigation was performed by simulating salt and pepper noise to these subsetted images. The process was carried out for noise density ranging from 0.1 to 0.9. The PSNR and MSE of the degraded images were determined. For image restoration, median filtering with different filter sizes was systematically applied. The size that can produce the restored image with the highest PSNR is chosen to be the optimal size for the particular noise density. This was based on the fact that the higher the PSNR, the higher the quality of the restored image and the lesser the remaining noise were left. The process was repeated for the rest of the images with other noise densities. The flowchart of the process is given in Fig. 9.



(a)

(b)

Fig. 7. Location of the Study Site: (a) Map of Malaysia and (b) Grand Hall of Universiti Teknikal Malaysia Melaka (UTeM).

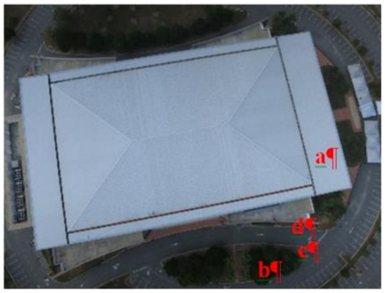


Fig. 8. UTeM's Grand Hall Captured from PRSS.

TABLE III. ATTRIBUTE AND THE CORRESPONDING METADATA OF THE IMAGERY

Image Attribute	Metadata
Date	27 March 2016
Location	2.31137, 102.3223
Altitude	110 m
Time	9.37 am
Image size	2.3 MB
Image dimensions	4000 x 3000
Shot info	1/2000 sec. f/5.6 5.2 mm
Iso	800
Device	Canon PowerShot S100
Platform	PRSS

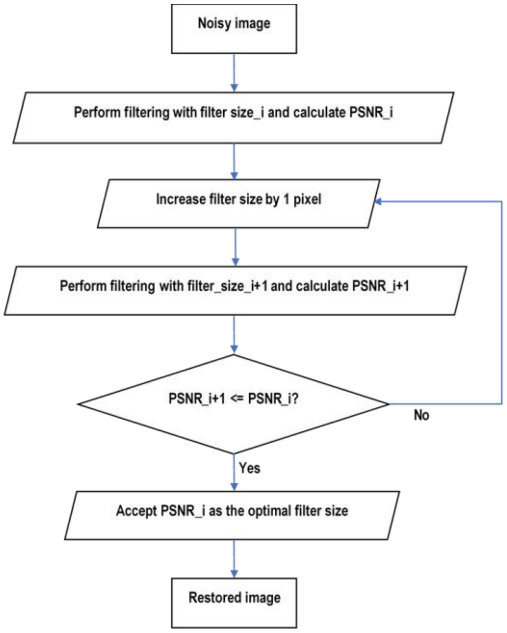

Noisy image
Perform filtering with filter size_i and calculate PSNR_i
Increase filter size by 1 pixel
Perform filtering with filter_size_i+1 and calculate PSNR_i+1
PSNR_i+1 <= PSNR_i?
No
Yes
Accept PSNR_i as the optimal filter size
Restored image

Fig. 9. Flowchart of the Image Restoration Process. Results and Discussion.

Fig. 10 shows the selected regions: (a) grand hall's roof surface, (b) landscape's vegetation, (c) road surface, and (d) balcony's roof surface that were subsetted from the grand hall image in Fig. 8. It is obvious that these regions possess different brightness conditions due to the different spectral properties of the materials [24]. Graphs of MSE versus noise density were then plotted to investigate the relationship between them for red, green and blue channel and for each of the regions.

Fig. 11 shows the MSE for (a) grand hall's roof surface, (b) landscape's vegetation, (c) road surface, and (d) balcony's roof surface. It can be seen that for all regions MSE increases as noise density increases. For grand hall's roof surface, the separation of the MSE for the red, green and blue channel is getting larger as noise density increases. Blue channel gives higher MSE compared to the green and blue channel for all noise densities. At 0.1 noise density, the MSE is approximately 2000 while at 0.9 noise density, the MSE ranging from 14000 to 20000 for all channels. A similar trend can be seen for road surface however with closer separation between the channels with 2000 and 16000 to 18000 MSE for 0.1 and 0.9 noise densities respectively. A different trend is shown for landscape's vegetation and balcony's roof surface where the curves are very close between each other with approximately 4000 to 25000 MSE and 2500 to 16000 MSE respectively at 0.1 and 0.9 noise density. Since MSE and PSNR are interrelated, graphs of PSNR versus noise density were then plotted to investigate the relationship between them for red, green and blue channel and for each of the regions.

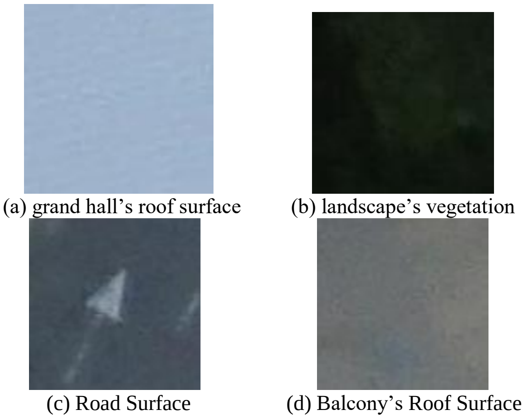
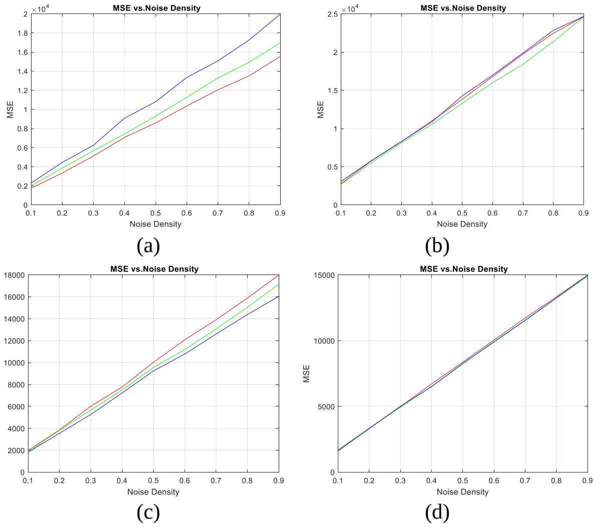

(a) grand hall's roof surface (b) landscape's vegetation
(c) Road Surface (d) Balcony's Roof Surface

Fig. 10. Selected regions: (a) grand hall's roof surface, (b) landscape's vegetation, (c) road surface and (d) balcony's roof surface.



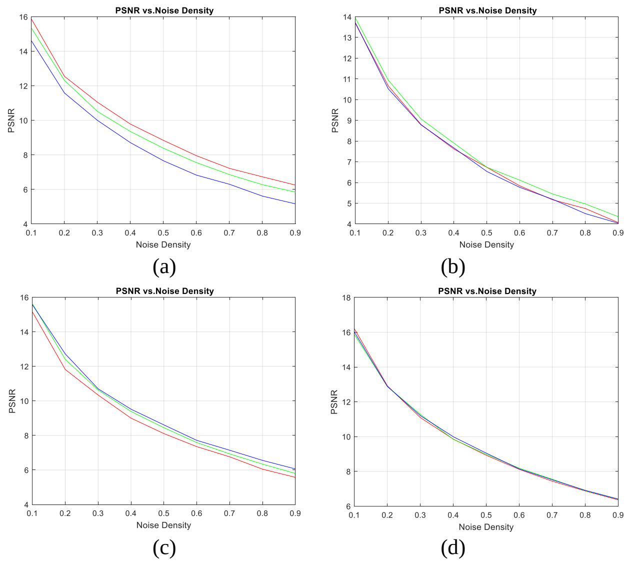
(a) (b)
(c) (d)

Fig. 11. MSE for (a) Grand Hall's Roof Surface, (b) Landscape's Vegetation, (c) Road Surface and (d) Balcony's Roof Surface.

Fig. 12 shows PSNR versus noise density for (a) grand hall's roof surface, (b) landscape's vegetation, (c) road surface and (d) balcony's roof surface. The colour of the curves indicates the PSNR for the red, green and blue channel of the image for each of the regions. For grand hall's roof surface, it is clear that for all channels PSNR decreases as noise density increases. At 0.1 noise density, for all channels, PSNR is below 16 while at 0.9 noise density, PSNR is above 4. There is an obvious separation between the PSNR curves with PSNR for the red channel seems higher than the green and blue channel. For landscape's vegetation, the PSNR curves are close between each other. At 0.1 noise density, the PSNR for all channels is below 14 while at 0.9 noise density, the PSNR for all channels is approximately 4. For the road surface, the separation of the PSNR curves is less than the grand hall's roof surface. At 0.1 noise density, the PSNR for all channels is below 16 while at 0.9 noise density, the PSNR for all channels is approximately 6. For balcony's roof surface, the PSNR curves are very close between each other. At 0.1 noise density, the PSNR for all channels is approximately 16 while at 0.9 noise density, the PSNR for all channels is above 6.

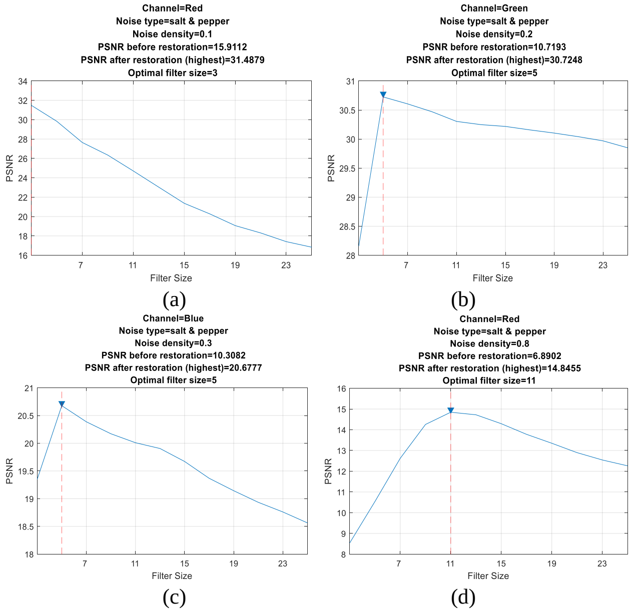
Previously the outcomes of the analysis in terms of MSE and PSNR for noisy images have been presented. Next, the restoration of these noisy images using median filtering was performed. In doing so, graphs of PSNR versus filter size were plotted for each noise density, for each channel and for each region. This allows the optimal filter size to be identified systematically. Following this, the variation of the optimal filter size as the noise density increases is analysed for each channel and for each region.

Fig. 13 shows PSNR versus filter size for grand hall's roof surface for selected noise density: (a) 0.1, (b) 0.2, (c) 0.3 and (d) 0.8 in red, green, blue and red respectively. The maximum PSNR is indicated by the highest peak marked with dotted red line and is taken to be the optimal filter size for the particular noise density and channel.



(a) (b) (c) (d)

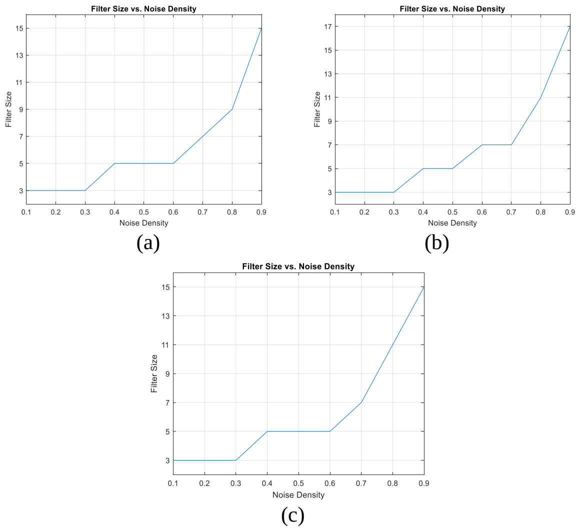
Fig. 12. PSNR for (a) Grand Hall's Roof Surface, (b) Landscape's Vegetation, (c) Road Surface and (d) Balcony's Roof Surface.



(a) (b) (c) (d)

Fig. 13. PSNR versus Filter Size for Grand Hall's Roof Surface for Selected Noise Density : (a) 0.1, (b) 0.2, (c) 0.3 and (d) 0.8 in Red, Green, Blue and Red Channel respectively.

Fig. 14 shows the optimal filter size for the grand hall's roof surface in (a) red, (b) green and (c) blue channel. For the red channel, the filter size varies from 3 to 15 with a gradual increase from 0.1 to 0.6 noise density while a rapid increase from 0.6 to 0.9 noise density. For the green channel, the filter size varies from 3 to 25 with a steady increase throughout the noise densities. For the blue channel, the trend is similar to the green channel. For the grand hall's roof surface, blue and green channel require filters with higher sizes compared to the red channel. This is due to the shorter wavelengths of the green and blue channels which experience a more significant degradation compared to the red channel with higher wavelengths, therefore require a higher filter size.

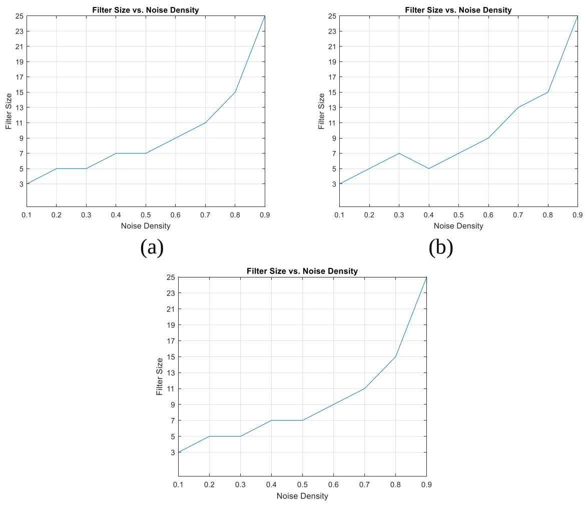


(a) (b) (c)

Fig. 14. Optimal Filter Size for Grand Hall's Roof Surface in: (a) Red, (b) Green and (c) Blue Channel.

Fig. 15 shows the optimal filter size for landscape's vegetation in red, (b) green and (c) blue channel. For all channels, there seems to be a steady increase in filter size from 0.1 to 0.9 noise density ranging from 3 to 25. However, there is a sudden drop in filter size at 0.4 noise density for the green channel. For landscape's vegetation, the degradation is not likely to be affected by the different wavelengths (correspond to different channels). This may be due to the very dark spectral properties of the landscape's vegetation that can somewhat compensate the effects of the salt and pepper noise and therefore require about similar filter size trend for all channels.

Fig. 16 shows the optimal filter size for road surface in (a) red, (b) green and (c) blue channel. For the red channel, the filter size varies from 3 to 19 with a slow increase in size from 0.1 to 0.6 noise density but a faster increase in size from 0.6 to 0.9 noise density. For the green channel, the filter size varies from 3 to 21 with a steady increase from 0.1 to 0.6 noise density but a more rapid increase from 0.6 to 0.9 noise density. Compared to the grand hall's roof and landscape's vegetation, here the filter size starts with 5 instead of 3, and it does not change until 0.4 noise density. This indicates that green channel is sensitive to low-density noise but the effects of the noise do not change much until the noise level is at about moderate level. The increase of filter size is at a constant rate from 0.6 to 0.8 noise density and the rate become much faster from 0.8 to 0.9 noise density in which indicating the effects of noise are more significant at moderate and much more significant at higher noise densities. For the blue channel, the filter size varies from 3 to 25 with a steady increase from 0.1 to 0.6 but a faster increase from 0.6 to 0.9. For the road surface, the shorter the wavelengths of the channels, the greater the effects of the salt and pepper noise, therefore the higher the filter size is required. This may be due to the moderately bright properties of the road surface that signifying the effects of salt and pepper noise.

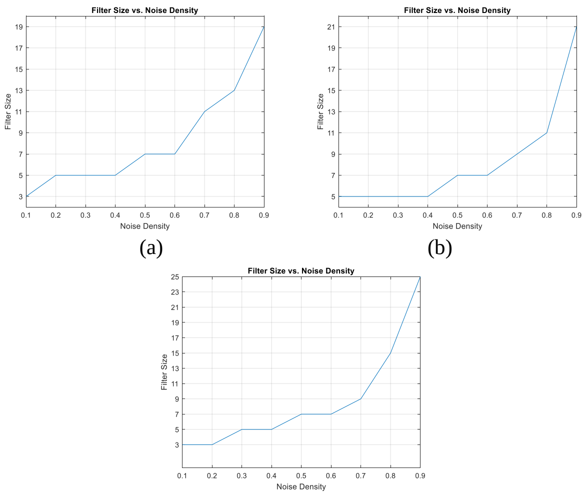


(a)

(b)

(c)

Fig. 15. Optimal Filter Size for Landscape's Vegetation in: (a) Red, (b) Green and (c) Blue Channel.



(a)

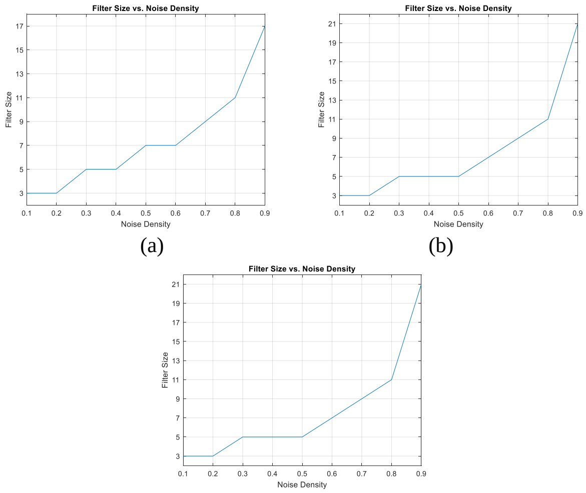
(b)

(c)

Fig. 16. Optimal Filter Size for Road Surface in: (a) Red, (b) Green and (c) Blue Channel.

Fig. 17 shows the optimal filter size for balcony's roof surface in (a) red, (b) green and (c) blue channel. For the red channel, the filter size varies from 3 to 17 with a steady increase from 0.1 to 0.6 noise density but a more rapid increase from 0.6 to 0.9 noise density. For the green channel, the filter size varies from 3 to 21, with a gradual increase from 0.1 to 0.5 but a rapid increase from 0.5 to 0.9 noise density. A similar trend is shown by the blue channel. For balcony's roof surface, the green and blue channel that having shorter wavelengths is being more affected by the salt and pepper noise and therefore require a higher optimal filter size compared to red channel that possesses longer wavelengths.

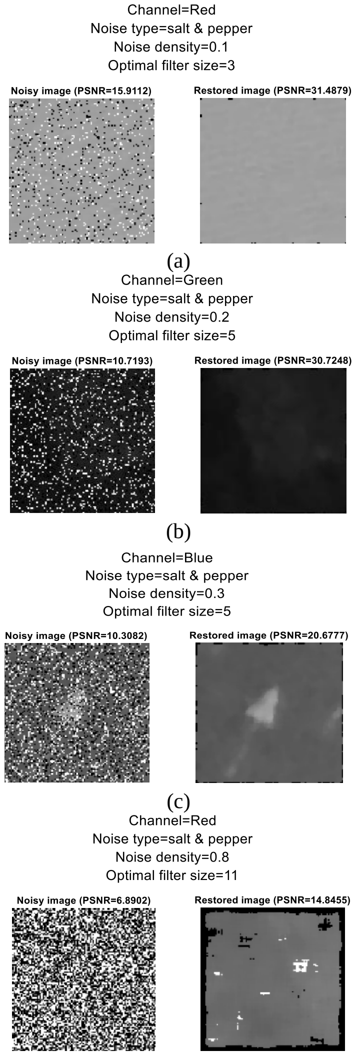
So far, the outcomes of the quantitative analyses have been presented and discussed. For the purpose of qualitative analysis, selected samples of images before and after restoration were displayed side by side so that visual inspection of the restoration performance can be validated. We purposely chosen three of images with low noise density (≤ 0.5) and one with high density (> 0.5) so that the outcomes are worth showing. Fig. 18 shows the noisy and restored images on selected samples for: (a) grand hall's roof surface, (b) landscape's vegetation, (c) road surface and (d) balcony's roof surface with noise density 0.2, 0.3, 0.5 and 0.8 respectively. It is clear that for low noise densities in (a), (b) and (c), the restoration works well where almost all noises are successfully removed. For (d), it can be seen that most noises are removed however there seems to be loss of information, indicated by bright and dark patches, within the balcony's roof surface image. The qualitative analysis shows that the restoration works best for images with low-density compared to high-density salt and pepper noises. To examine the overall filter size variation trend for all regions and channels, the minimum, maximum and variation of the optimal filter sizes were tabulated in a single table.



(a) (b)

(c)

Fig. 17. Optimal Filter Size for Balcony's Roof Surface in: (a) Red, (b) Green and (c) Blue Channel



(a)

(b)

(c)

(d)

Fig. 18. Noisy and Restored Images on Selected Samples for: (a) Grand Hall's Roof Surface, (b) Landscape's Vegetation, (c) Road Surface, and (d) Balcony's Roof Surface.

Table IV shows the minimum, maximum and variation for the filter size for the red, green and blue channel based on region A, B, C and D representing grand hall's roof surface, landscape's vegetation, road surfaceand balcony's roof surface respectively. R_{min}, G_{min} and B_{min} are minimum filter sizes in red, green and blue channel respectively. R_{max}, G_{max} and B_{max} are maximum filter sizes in red, green and blue channel respectively. $R_{var.}$, $G_{var.}$ and $B_{var.}$ are the variation of filter sizes in red, green and blue channel respectively, avrg. is the average value of respective components while var. avrg. is the average of $R_{var.}$, $G_{var.}$ and $B_{var.}$.

TABLE IV. MINIMUM, MAXIMUM AND VARIATION FOR FILTER SIZE FOR THE RED, GREEN AND BLUE CHANNEL BASED ON REGION. A, B, C AND D REPRESENTING GRAND HALL'S ROOF SURFACE, LANDSCAPE'S VEGETATION, ROAD SURFACE AND BALCONY'S ROOF SURFACE RESPECTIVELY

Region	R_{m} in	R_{m} ax	R_{v} ar.	G_{m} in	G_{m} ax	G_{v} ar.	B_{m} in	B_{m} ax	B_{va} r.	var. avrg. $\frac{(R_{var.}+G_{var.}+B_{v}}{3}$
A	3	15	12	3	17	14	3	15	12	12.7
B	3	25	22	3	25	22	3	25	22	**22**
C	3	19	16	5	21	16	3	25	22	18
D	3	17	14	3	21	18	3	21	18	16.7
avrg.	3	19	16	3.5	21	17.5	3	21.5	**18.5**	

It can be seen that the average value of R_{min} and B_{min} is 3 and is the smallest while the average of B_{max} is 21.5 and is the largest. The variation between the average of B_{min} and B_{max}, $B_{var.}$ is 18.5 for which is the highest. This indicates the filter for blue channel easily changes as noise density changes. Thus the blue channel is the most sensitive to noise compared to the red and green channel. This is because the blue channel has a higher ability to capture noise effects compared to the red and green channel, hence noise in the blue channel has a higher visibility compared to the red and green channel. In term of average filter variation, landscape's vegetation has the highest variation of 22 signifying that the very dark spectral properties of landscape's vegetation are easily being influenced by the effects of noise, thus has the highest sensitivity to noise compared to brighter regions

V. CONCLUSION

In this study, we have experimented salt and pepper noise of different densities on the red, green and blue channel of a UAV image containing regions with different spectral properties. Image restoration has been performed using median filtering of different filter sizes. An optimal filter size has been chosen based on the highest PSNR of the restored image produced. The result shows that the effects of noise on a UAV image and the optimal size of a median filter for image restoration are dependent on the spectral properties of the channels and regions of interest. The blue channel is found to have the highest response to noise due to the shortest spectral wavelengths compared to the red and green channel, while landscape's vegetation is the most sensitive to noise compared to grand hall's roof surface, road surface and balcony's roof surface due to its very dark spectral properties that making it easily being influenced by the noise effects. For image restoration, generally, optimal median filter size increases with different rate of variation as noise density increases. The restoration works best for images with low-density compared to high-density salt and pepper noises. The filter size for blue channel varies with the biggest variation and is the largest for the highest noise density due to the higher response to noise effects compared to the red and green channel. Darker regions require larger filter sizes compared to brighter regions as noise density increases due to the higher sensitivity to the presence of noise.

Acknowledgment

The authors thank Universiti Teknikal Malaysia Melaka (UTeM) for funding this research under the Ministry of Higher Education grant FRGS/2/2014/ICT02/UTEM/02/2.

References

[1] T. Fellmann, P. Witzke, and F. Weiss, "Major challenges of integrating agriculture into climate change mitigation policy frameworks", Mitig Adapt Strateg Glob Change, vol. 23, pp. 451–468, 2018.

[2] A. Ahmad and S. Quegan, "Multitemporal cloud detection and masking using MODIS data", Applied Mathematical Sciences, vol. 8, no. 7, pp. 345-353, 2014.

[3] S. P. A. R. Putra, S. C. Keat, K. Abdullah, L. H. San and M. N. M. Nordin, "Texture analysis of AIRSAR images for land cover classification", Proceeding of the 2011 IEEE International Conference on Space Science and Communication (IconSpace), Penang, pp. 243 – 248, 2010.

[4] B. Bansod, R. Singh, R. Thakur and G. Singhal, "A comparison between satellite based and drone based remote sensing technology to achieve sustainable development: a review", Journal of Agriculture and Environment for International Development, vol. 111, no. 2, pp. 383 – 407, 2017.

[5] N. A. Sari, A. Ahmad, M. Y. A. Sari, S. Sahib and A. W. Rasib, "Development of rapid low-cost LARS platform for oil palm plantation", Jurnal Teknologi, vol. 77, no. 20, pp. 99 – 105, 2015.

[6] R. Wang, X. Xiao, B. Guo, Q. Qin and R. Chen, "An effective image denoising method for UAV images via improved generative adversarial networks, Sensors", vol. 18, no. 1985, pp. 1 – 23, 2018.

[7] A. Ahmad and S., Quegan, "The effects of haze on the spectral and statistical properties of land cover classification", Applied Mathematical Sciences, vol. 8, no. 180, pp. 9001 - 9013, 2014.

[8] N. Mesner and K. Ostir, "Investigating the impact of spatial and spectral resolution of satellite images on segmentation quality", J. Appl. Rem. Sens., vol. 8, 083696-1 – 083696-14, 2014.

[9] A. Matese, P. Toscano, S. F. Di Gennaro, L. Genesio, F. P. Vaccari, J. Primicerio, C. Belli, A. Zaldei, R. Bianconi and B. Gioli, "Intercomparison of UAV, aircraft and satellite remote sensing platforms for precision viticulture", Remote Sens., vol. 7, pp. 2971-2990, 2015.

[10] Google Maps, "Dewan Canselor UTeM", 1:1000. Google Maps [online] Available through: UTeM FTMK, 2020. [Accessed 23 October 2020].

[11] A. Ahmad, S. Quegan, S. F. Sufahani, H. Sakidin and M. M. Abdullah, "Haze effects on satellite remote sensing imagery and their corrections", International Journal of Advanced Computer Science and Applications (IJACSA), vol. 10, no. 10, pp. 69 – 76, 2019.

[12] D. C. Tsouros, S. Bibi and P. G Sarigiannidis, "A review on UAV-based applications for precision agriculture", Information, vol. 10, no. 349, pp. 1 – 26, 2019.

[13] R. C. Gonzalez and R. E. Woods, "Digital Image Processing", 3rd Edition, Pearson Education Inc., 2008.

[14] G. Kaur, R. Kumar and K. Kainth, "A review paper on different noise types and digital image processing", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6, no. 6, pp. 562 – 565, 2016.

[15] Medium. Noise in Digital Image Processing, [online] Available at: <https://medium.com/image-vision/noise-in-digital-image-processing-55357c9fab71>, 2020, [Accessed 12 October 2020].

[16] L. Tan and J. Jiang, "Digital Signal Processing: Fundamentals and Applications", 3rd Edition, Academic Press, 2019.

[17] A. Ahmad, M. K. A. Ghani, S. Razali, H. Sakidin and N. M. Hashim, "Haze reduction from remotely sensed data", Applied Mathematical Sciences, vol. 8, no. 36, pp. 1755 - 1762, 2014.

[18] A. Ahmad and S. Quegan, "Analysis of Maximum Likelihood classification technique on Landsat 5 TM satellite data of tropical land covers", Proceedings - 2012 IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2012, pp. 280 – 285, 2012.

[19] Purwadi, N. Suryana, N. A. Abu and B. A. Kusuma, "A comprehensive review: Classification techniques on hyperspectral remote sensing", International Journal of Advanced Trends in Computer Science and Engineering, vol. 8, no. 1.5, pp. 156 – 164, 2019.

[20] S. Pisupati and M. B. Ismail, "Image registration method for satellite image sensing using feature based techniques", International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 1, pp. 590 – 593, 2020.

[21] S. Semenov, D. Voloshyn and A. N. Ahmed, "Mathematical model of the implementation process of flight task of unmanned aerial vehicle in the conditions of external impact", International Journal of Advanced Trends in Computer Science and Engineering, vol. 8, no. 1.2, pp. 7 – 13, 2019.

[22] M. Y. A. Sari, A. Ahmad, Y. M. M. Hassim, S. Sahib, N. A. Sari and A. W. Rasib, "Large scale topographic map comparison using unmanned aerial vehicle (UAV) imagers and real time kinematic (RTK)", International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 1.1, pp. 328 – 338, 2020.

[23] M. Y. A. Sari, A. W. Rasib, H. M. Ali, A. R. M. Yusoff, M. I. Hassan, K. M. Idris, A, Ahmad and R. Dollah, "3D mapping based-on integration of uav platform and ground surveying" International Journal of Advanced Computer Science and Applications (IJACSA), vol. 9, no. 12, pp. 160 - 168, 2018.

[24] A. Ahmad, U. K. M. Hashim, O. Mohd, M. M. Abdullah, H. Sakidin, A. W. Rasib and S. F. Sufahani, "Comparative analysis of support vector machine, maximum likelihood and neural network classification on multispectral remote sensing data " International Journal of Advanced Computer Science and Applications (IJACSA), vol. 9, no. 9, pp. 529 – 537, 2018.

Prevention of Attacks in Mobile Ad Hoc Network using African Buffalo Monitoring Zone Protocol

R.Srilakshmi[1]
Research Scholar
Koneru Lakshmaiah Education Foundation,
Vaddeswaram, 522502, Andhra pradesh, India

Dr. M.Jaya Bhaskar[2]
Professor
Koneru Lakshmaiah Education Foundation,
Vaddeswaram, 522502, Andhra Pradesh, India

***Abstract*—Mobile ad hoc networks (MANET) can be utilized for communicating wirelessly. However, MANET is affected by many attacks and malicious activities. In MANET, the prevention approach is necessary to secure communication. MANET is easily affected by numerous attacks such as wormhole (WH) attack, Grey-hole (GH) attack, and black-hole (BH) attack in which the sender hubs can't able to transmits the message to the target node due to the malicious behavior. To prevent the attacks in MANET, this research introduces a novel routing protocol as African Buffalo Monitoring Zone Protocol (ABMZP). This approach is utilized for preventing wormhole attack and other malicious activities in MANET. This mechanism monitors the communication channel continuously and identifies the attack detection. Sequentially, the ABMZP approach prevents the harmful nodes and finds the alternate path for communication. The simulation of this research is done with the use of Network Simulator 2 (NS-2) and finally, the efficiency of the projected ABMZP work outcomes are compared with the latest existing techniques and provides superior results.**

Keywords*—Mobile ad hoc network; malicious nodes; routing protocol; wormhole attack; security

I. Introduction

In general, MANET referred to the decentralized classification of a wireless network [1, 44]. Also, MANET has a routable associating situation on upper of a connection layer [2, 35]. Also, MANET has a collection of nodes, which are communicating wirelessly [3, 37]. Moreover, the nodes in the MANET are freely moving as the system topology transforms frequently [4,]. Generally, every hub acts as a router when it sends traffic to another specific node in the network. The MANET nature is very dynamic that is utilized for communicating between two nodes during business conferences, natural disasters, etc [5, 41]. In MANET, the routing procedure is utilized for communicating between the pair of nodes. These nodes are having the ability to transfer the message from the source hub to the target node [6, 40]. Generally, MANET has several groups of nodes, which is no permanent infrastructure to connect the nodes. So, these are very flexible & smoothly reconfigurable and these networks required a limited number of properties like memory of the network, bandwidth, and battery & computation power [7].The MANET nodes are in a particular network range and communicate directly to each other [8]. In general, several network actions are achieved using mobile nodes in MANET such as packet forwarding, packet detection, packet communication, and network organization [9].

Moreover, MANET is having the features of communicating in free space, broad sharing of packets, and nodes. However, MANET is vulnerable to several types of malicious activities so prevention is necessary to protect the channel [10, 39]. The packets should be controlled because when the sender node sends the message the neighboring nodes act maliciously to drop the data are passed through it [11, 38]. Several attacks are affecting the MANET network such as BH, denial of service (DoS), WH, distributed DoS(DDoS) and GH attacks. Moreover, the difficulties of misconduct routing are one of the disseminated protections terrorizations in the network like BH attacks. Therefore, several investigators are proposing many protected routing ideas to overcome these issues, but the safety problems of network are still an issue.

So, several prevention mechanisms are introduced to prevent malicious attacks but, it has attained numerous challenges. Several routing protocols like artificial intelligence [12], EMAODV [13, 45], and Ad-hoc On-demand Distance Vector (AODV) mechanism are introduced to secure communication. However, the malicious activities create trustful nodes so the sender transmits the message to it [14, 36]. But, it can't able to reach the destination because the hacker transmits the message into the third person [15, 46]. So, the proposed approach introduces a novel protocol for avoiding the occurrences in the MANET and provides secure communication. The proposed mechanism improves network performance and provides high security.

This research is categorized as various sections that are Section 2 demonstrates the recent literatures about the security, the problem definition is given in Section 3; the proposed methodology is detailed in Section 4, the attained results are explained in Section 5, and the conclusion part is mentioned in Section 6.

II. Related Work

Various authors have given different approaches to eliminate attacks in MANET and here are some:

A BH attack is one of the categories of MANET occurrence and here the malevolent hubs become part of the network. As a result, it absorbs everything that comes in, because it acts as a hole and the packets fall into it. Therefore, the desired target node does not attain the data packets; therefore, it disturbs the entire communication. Here, Gupta *et al* [16] established the methodology using reliability factors for detecting and preventing the BH attacks in MANET. Also, the AODV protocol is modified to secure the channel that can act against the BH attack. Thus, the fake RREQ conception is utilized for detecting the malicious nodes in this method. Furthermore, this approach decreases the amount of dropped packets.

In MANET, several intrusions are detected to affect the nodes in the network. According to this reason, Su and Ming-Yang [17] introduced Anti-Blackhole Mechanism (ABM) for reducing the BH attack in the MANET. Thus, it detects and separates the malicious nodes in the MANET. Here, the utilized IDS nodes are arranged in the form of sniff for achieving the function of ABM that is utilized for evaluating the uncommon value of a node. Here, it considered the threshold value and when the particular node threshold is increased then IDS is blocking the message.

Moreover, WH attack is the rigorous activities in MANET, which is defeated using many prevention mechanisms. These prevention approaches are based on packet traversal time, round trip time, & hop-count but, these solutions are not successful. To avoid these issues, Vo *et al* [18] introduced a multi-level authentication method and procedure (MLAMAN). This method permitted the nodes to validate the packets based on 3 steps that are packet level, membership, and neighborhood level. Hence, the MLAMAN approach detects and prevents the wormhole attack in the network.

In MANET, the messages are not reaching the target node because of the malevolent activities of the node such as DoS and black-hole attacks. The packets are dropped because of the BH attacks that are eliminated by a new approach. Here, Gurung*et al* [19] introduced an approach that depends on a dynamic threshold algorithm that is named as MBDP-AODV protocol. This protocol moderates the collision of attacks based on various network densities and this mechanism improved network performance like throughput, PDR, overhead, and decreases the routing load.

MANET can be affected by many malicious attacks like DDoS attacks. A kind of DDoS attack is a Jellyfish attack that is quite hard and affects the complete performance of the network. To overcome this jellyfish attack, Doss *et al* [20] proposed an attack prevention APD-JFAD approach. This approach selects the trusted nodes for creating a path to overcome the attack. Moreover, it acts against the jellyfish attack accurately and provides better performance. Thus, the existing techniques presentation is summarized in the Table I.

The key contributions of the research are summarized below:

- In general, MANET is vulnerable to malicious activities. So, this approach develops the novel protocol for predicting malicious activities in the network.
- Here, African Buffalo Monitoring Zone Protocol (ABMZP) is developed to prevent the network from the WH, BH, and GH attacks.
- It secures the MANET communication channel against the malicious attacks.
- The implementation of the proposed approach is done using the NS-2 tool.
- Finally, the attained implementation results prove the efficacy of the projected approach.

TABLE I. RECENT LITERATURE BASED ON PREVENTING ATTACKS IN MANET

Author	**Method**	**Merits**	**Demerits**
Gupta *et al* [16]	Reliability Factor Based AODV Protocol (RF based AODV)	Packet drop ratio is reduced in this approach.	RF based AODV attained high simulation time and low accuracy.
Su and Ming-Yang [17]	Anti-Black hole Mechanism (ABM)	In this ABM method Packet loss rate is decreased.	It has high overhead.
Vo *et al* [18]	multi-level authentication model and protocol (MLAMAN)	It can detect the malicious activities as different tunnel lengths and node speeds.	This MLAMAN strategy attained high overhead during data transmission.
Gurung*et al* [19]	MBDP-AODV	MBDP-AODV manner attained high PDR, and throughput.	It attained high routing overhead during transmission.
Doss *et al* [20]	APD-JFAD	It attained high PDR, throughput, and low delay.	APD-JFAD method prevention accuracy is low.

III. PROBLEM STATEMENT

MANET has multiple nodes connected to multiple wireless connections. MANET is affected by various limitations such as limited bandwidth, connection failure, limited communication power, and power outage. Moreover, the wireless link and nodes are highly susceptible to attacks. The communication channel is affected by several malicious activities such as WH attacks [21], GH attacks [22], DDoS attacks [23], and BH attacks [24]. The most common attack in MANET is the WH attack, which can be dropped the packets and break the link. So, the information was not able to attain the destination of the network [25, 31]. Moreover, these attacks break the wireless links and pass the data into unwanted nodes. The destination node can't receive the packets because of the malicious activities [26, 32].

The source hubs transmit the data packs to the target through neighboring nodes. The attacker node receives the data packets from the sender. After that, the attacker node drops the data packets, which are detailed in fig.1. So, MANET security is necessary and that should be complete the security parameters such as network overload, processing time, and energy consumption. The proposed approach provides better network performance and high security between nodes.

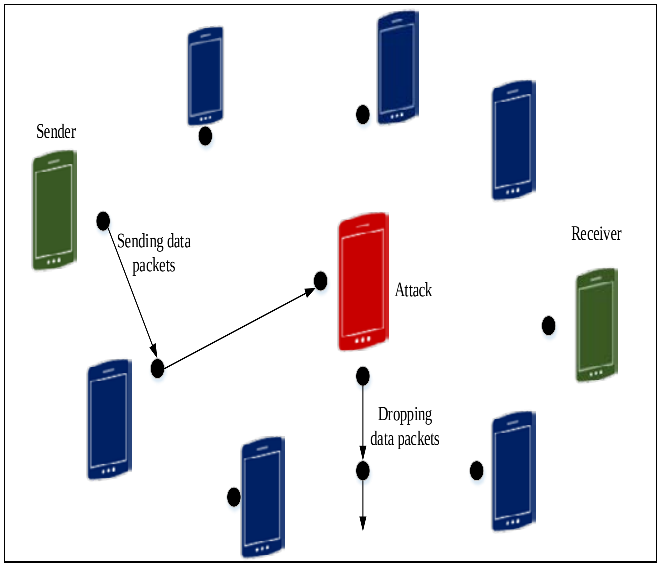


Fig. 1. System Framework for attack Intrusion in MANET.

IV. PROPOSED (ABMZP) METHODOLOGY

The proposed approach provides secure communication between the nodes in MANET. Initially, a communication channel is formed in MANET that can be easily affected by many attacks like WH, BH, and GH attacks. So, this methodology introduces an innovative African buffalo Monitoring zone protocol (ABMZP) to prevent communication channels from malicious activities. Finally, this approach acts as a prevention mechanism against the harmful nodes in the network. The proposed manner is aimed to secure the network before any type of malicious activity enters the node and make the malicious node not able to break the security of the network, which are shown in Fig. 2.

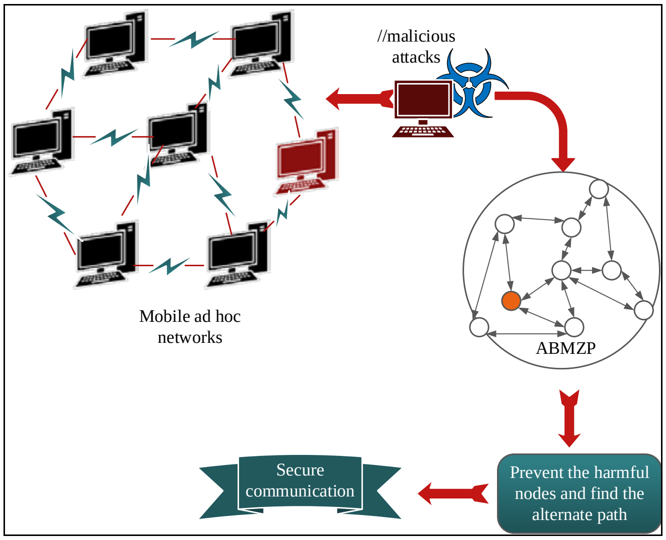


Fig. 2. Proposed Methodology.

A. *African Buffalo Monitoring Zone Protocol (ABMZP)*

African buffalo Monitoring Zone protocol (ABMZP) is the hybrid form of African buffalo optimization (ABO) [27,34] and Zone protocol (ZP) [28,43]. In this approach, the malicious nodes are predicted using the fitness function of ABO and it identifies another secure path for communication. Initially, the ABMZ protocol creates the routing zone for data transmission. This routing zone has several numbers of nodes (N) that all are in the communication network. Initially, the source hub conveys the data packets to the receiver through neighboring nodes that makes the path for packet transmission. MANET is vulnerable to malicious activities so the proposed approach identifies the attacked nodes and provides better transmission through alternate path [33,42]. The proposed ABMZP model identifies the attacks like WH attack, BH attack and GH attack. This approach initializes the network zone (N) that has several nodes. Primarily, the $R\,\mathrm{Re}\,q$ message is transmitted for all neighboring nodes by the sender node and that routers send the $RRply$ message to the sender node. Here, the proposed ABMZP monitors the network to identify the best path for better communication. Also, the proposed ABMZP categorizes the IP address for every nodes in the network. The finest path identification for better communication using eq. (1):

$$k'_w = P_k + N\left(\ln_1(bN - Q_k) + (\ln_2(bM.k. - Q_k))\right) \tag{1}$$

Where, N specifies the whole communication network, k denotes the communication path, P_k represents all nodes in the network, Q_k is the malicious nodes, $\ln_1$ and $\ln_2$ is the learning parameters, bN mentioned the IP address of all nodes and bM denotes the IP address of malicious nodes.

Algorithm 1: ABMZP for attack prevention

```
begin
{
 Initialize the communication network (N)
 Develop the source (S) and destination (D) nodes;
 Randomly initialize network path $K_w$ on the search space
Let the attack nodes as ($Q_k$) // $Q_k$ involves WH, BH, and GH attacks
 $P_k \rightarrow IP\_(N) = (bN)$ // set IP address for all nodes in the network
If $S \xrightarrow{R\,\mathrm{Re}\,q} P_k$ then $P_k$ & $Q_k \xrightarrow{RRply} S$
Analyze the network path using AB function by eqn.(1)
If $P_k(bN) \neq Q_k(bM)$ //bN and bM denotes the IP address of trusted node & Malicious node
then
Calculate the energy levels of all nodes using eqn.(2)
If $E_T(N) > 1J$ // malicious activities are present
then
      calculate the energy threshold for all nodes
      $1J < E_T(w) > 1.5J$ // $E_T(w)$ denotes the energy threshold of wormhole attack
      $1.5J < E_T(b) > 2J$ // $E_T(b)$ denotes the energy threshold of black hole attack
      $2J < E_T(g) > 2.5J$ // $E_T(g)$ denotes the energy threshold of grey-hole attack
end
alert the source node;// S node not transmits the message in this path
Identify the optimal path // secure communication
end-if
}
stop
```

Subsequently, if ABMZP detect the harmful nodes then it detects the attacks like WH, BH, and GH based on the energy levels of malicious nodes. Thus, the energy threshold of the nodes are identified using eq. (2).

$$E_T(N) = \frac{E_{R\,\mathrm{Re}\,q} + E_{RRply}}{Total\_energy} \qquad (2)$$

Where, $E_{R\,\mathrm{Re}\,q}$ is represents the energy for $R\,\mathrm{Re}\,q$ message, E_{RRply} is denotes the energy for $RRply$ message and λ is the energy of attacks. The process of ABMZP is explained using algorithm 1.

In this approach, the ABMZP monitors the network continuously to detect the harmful nodes. If it is identifying the harmful nodes then it alerts the source node and provides secure communication through finest path that process is explained in Fig. 3.

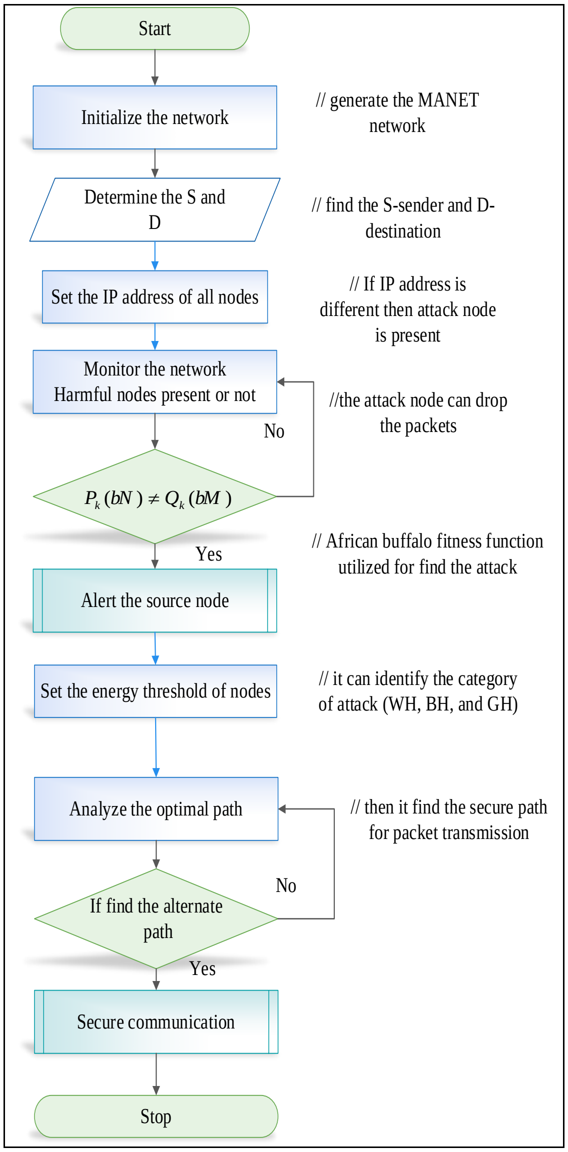
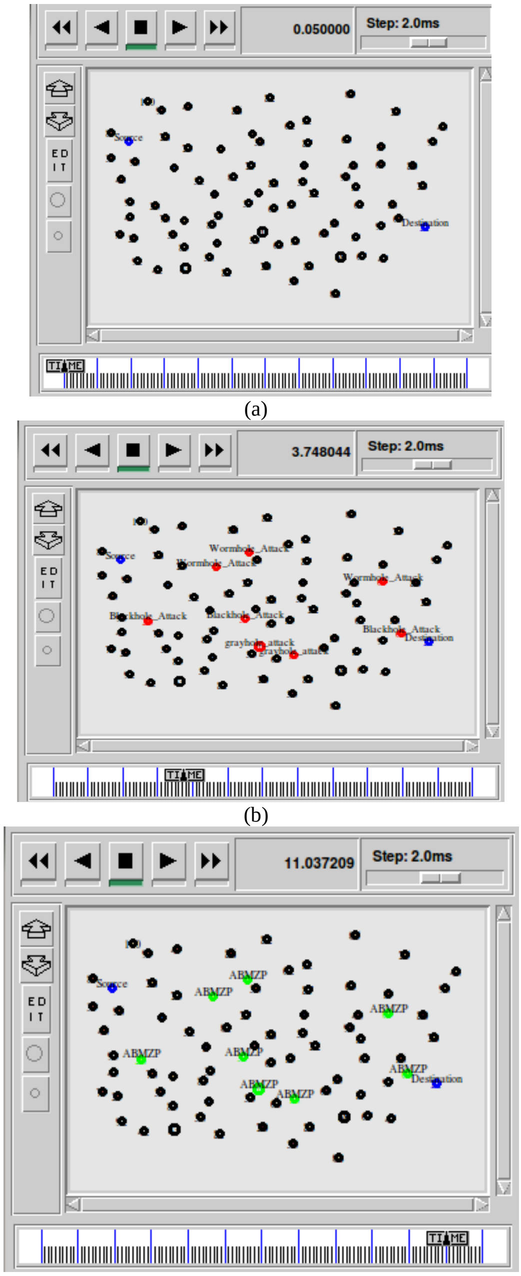

Start
Initialize the network
// generate the MANET network
Determine the S and D
// find the S-sender and D-destination
Set the IP address of all nodes
// If IP address is different then attack node is present
Monitor the network Harmful nodes present or not
//the attack node can drop the packets
No
$P_k(bN) \neq Q_k(bM)$
Yes
// African buffalo fitness function utilized for find the attack
Alert the source node
Set the energy threshold of nodes
// it can identify the category of attack (WH, BH, and GH)
Analyze the optimal path
// then it find the secure path for packet transmission
No
If find the alternate path
Yes
Secure communication
Stop

Fig. 3. Flow Chart for ABMZP.

V. Results and Discussion

The developed ABMZP model is simulated byNS-2 running on NAM console v1 15 in the Ubuntu 12.04.5 LTS platform. Generally, MANET is vulnerable to attacks so this proposed approach introduced novel prevention mechanism ABMZP. The projected ABMZP model is utilized to create the MANET nodes and transmit the messages through the routers in a secure way. This research focuses on the MANET nodes and identifies the malicious nodes. Also, ABMZP creates a better channel for communication, which is detailed in Fig. 4. Finally, the performance metrics are calculated using NS-2 and provide better network performance compared with existing approaches.



(a)

(b)

(c)

Fig. 4. (a) Node Formation in Network, (b) Detection of Attacks, (c) Communication through Best Path.

B. Case Study

Let the sender node S, destination node D, and A, B, C, D, E, F,G,H,I,J, and Kare taken as other neighboring nodes. Here, the data packets are transferred from S to D through neighboring nodes. Initially, the sender node wants to convey the information to the target. So, the sender node searches the IP address of every nearby node. Moreover, the sender node finds the IP address of each adjoining node in the zone. Subsequently, every node has a dissimilar IP address and that is stored in the sender node. Primarily, the sender node transmits the Route Request (RReq) to the routers in the zone before transmitting the packets. Consequently, the neighboring nodes are send Route Reply (RRply) message to the sender, if any node is attacked by malicious that is also sending the RRply message. The ABMZP is always monitoring the network and RRply messages. Also, it analyzes the IP address of the nodes which are forward the message to the sender. If it is same means there no malicious activity and it is different means the attack present in the network. The graphical illustration of these details is represented in Fig. 5.

Furthermore, malicious node forwards the RRply message through the sender, which is detected using ABMZP. Let us consider the total energy of the nodes as 0.4J and the energy for RReq message as 0.2J and the energy for RRply message as 0.18J, which are substitute in eqn.2. So, the attained energy threshold for all nodes as E(N)=0.875J. Also, let the energy threshold level for attacks as WH attack Ew= 1.2J, BH attack Eb=1.6J, and GH attack Eg=2.1J. Here, the proposed ABMZP utilizes lower energy to transmit the packets. If the energy threshold is high than the particular range then it is affected with malevolent activities. So, the ABMZP protocol alerts the sender node and transmits the message in another path to provide secure transmission, which is shown in Fig. 6.

Consequently, the proposed ABMZP develops best path to secure transmission. Finally, the data packets are sending through the secure path and reach the destination.

C. Performance Metrics

The introduced ABMZP mechanism compute the parameters like throughput, overhead, packet delivery ratio (PDR), packet delay, and end_to_end delay. This manner is compared with other strategies such as MBDP-AODV [19], OLSR [29], and CLPDM-SI [30]. The proposed method achieved better results in overhead, packet delay, PDR, and throughput.

1) Attack prevention rate: This is represented as a metrics for calculating the effectiveness of attack prevention rate in the network. It is the fraction of total quantity of sends messages and sum of data packets attained to the destination and the mathematical expression is represented in Eq. (3):

$$A = \frac{(Tn' + Tp')}{(Tn' + Tp' + Fn' + Fp')} \tag{3}$$

Where, Tn' is True negative, Tp' is denoted true positive, Fn' is represented the false negative value and Fp' is symbolized the false positive.

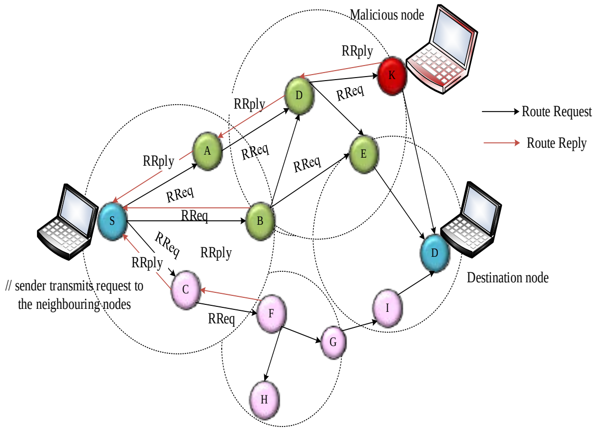


Fig. 5. Network Zone for Data Transmission.

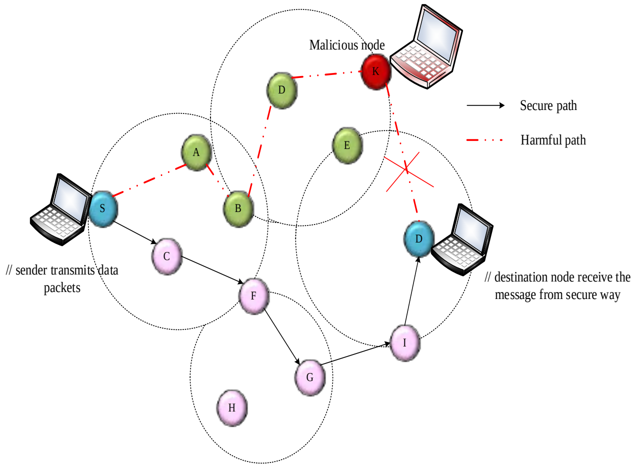

Malicious node
K
D
Secure path
E
Harmful path
A
S
B
D
C
// sender transmits data packets
F
// destination node receive the message from secure way
I
G
H

Fig. 6. Secure Packet Transmission.

TABLE II. CALCULATION OF ATTACK PREVENTION RATE

Method	Attack Prevention rate (%)
MBDP-AODV[19]	89
OLSR [29]	98.5
CLPDM-SI [30]	96
Proposed [ABMZP]	99. 96

The attack prevention rate is proves the efficiency of the proposed ABMZP method. This is campared with some techniques like MBDP-AODV, OLSR, and CLPDM-SI. Here, MBDP-AODV attains lower prevention rate as 89%, OLSR, and CLPDM-SI attains 98.5% & 96% prevention rates. But, the proposed ABMZP attains 99.96 % high rate for attack preventing rate, these values are given in Table II and it is represented in Fig. 7.

2) PDR calculation: It is the ratio between the entire amount of attained data packets and quantity of sent packets, which is calculated using eq. (4). The PDR ratio of the proposed ABMZP model is detailed in Fig. 8.

$$PDR = \frac{No.of\_received\_packets}{No.of\_sent\_packets} X100 \tag{4}$$

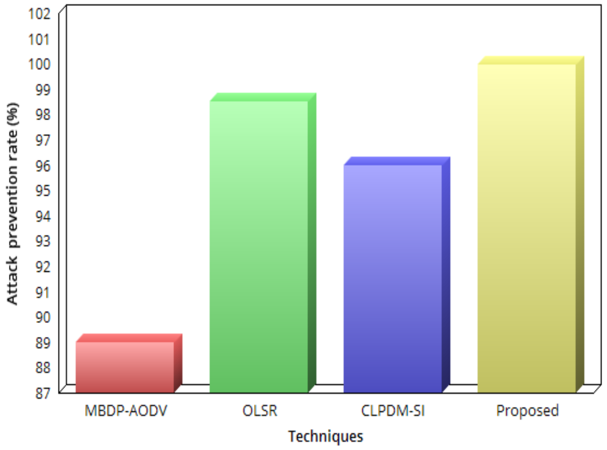


Fig. 7. Evaluation of Attack Prevention Rate.

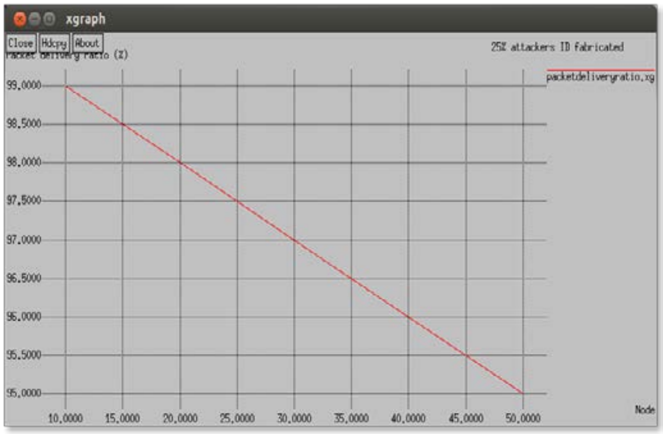


Fig. 8. PDR Rate.

Moreover, the PDR ratio is calculated and evaluated using prevailing approaches. Here, the PDR ratio is computed based on the number of transmitted nodes. While considering 50 numbers of nodes, MBDP- AODV achieved 92% PDR, OLSR attained 95% PDR and CLPDM-SI attained 99% PDR. Moreover, the proposed method ABMZP achieved 99.32% high PDR rate, which is given in Table III and represented in Fig. 9.

3) End-To-End delay calculation: It is a calculation of the regular time when the data take a time period to reach the target node from the sender and the delay time is calculated using eq. (5).

$$End\_to\_end-delay=\frac{received\_packets\_time}{sent\_packets\_time}$$

Generally, high packet delay increased the number of retransmitted RReq messages and data packets. Also, it can easily reduce the network resources and wastes the energy of nodes. In this approach, the delay is very low because of high attack prevention, which is mentioned in Fig. 10.

This ABMZP outcome is compared with MBDP-AODV, OLSR, and CLPDM-SI. Here, the MBDP-AODV method has a high delay for transmitting the packets. Therefore, the novel ABMZP method provides a low delay for receiving the packets in the destination that is given in Table IV and represented in Fig. 11.

TABLE III. EVALUATION OF PDR

Nodes	**Packet Delivery ratio (%)**			
	MBDP AODV[19]	**OLSR [29]**	**CLPDM-SI [30]**	**Proposed [ABMZP]**
10	94	48	85	99.98
20	93	78	98	99.86
30	92	82	99	99.65
40	90	88	98	99.55
50	92	95	99	99.32

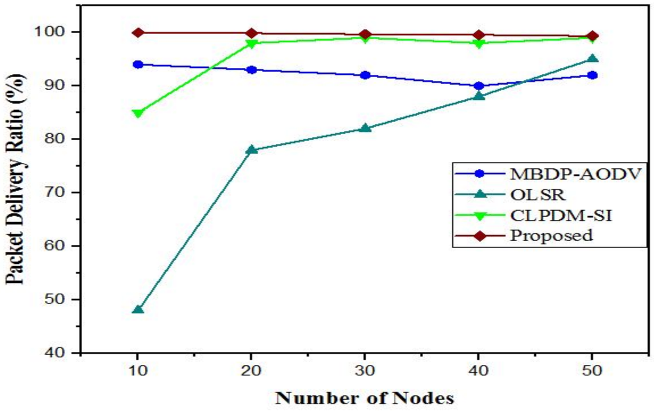


Fig. 9. Comparison of PDR Ratio.

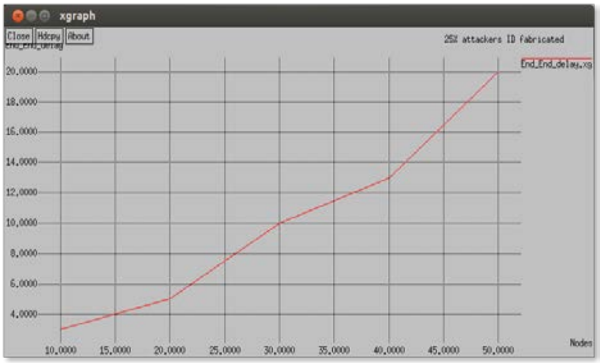


Fig. 10. End-End- Delay Calculation.

TABLE IV. EVALUATION OF END-TO-END DELAY

Nodes	**End to end delay (s)**			
	MBDP AODV[19]	**OLSR [29]**	**CLPDM-SI [30]**	**Proposed [ABMZP]**
10	40	18	8	3
20	80	28	10	5
30	110	38	14	10
40	160	58	28	13
50	180	64	30	20

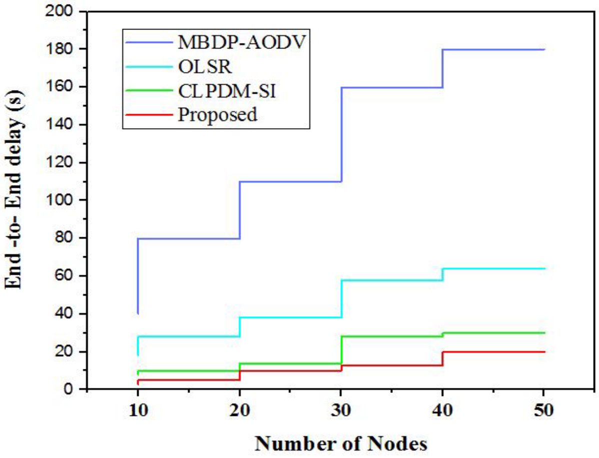


Fig. 11. Comparison of End-To-End Delay.

The security mechanism is taken at a particular time for detecting and eliminating the harmful nodes. Thus, the MANET security mechanisms should have very low processing time for packet transmission.

4) Packet drop ratio: This calculation measured using the fraction of total packet loss during transmission to the entire amount of received data packets. The sender hub transmits the packets to the neighboring nodes. If any attack present in the network then it is dropped or losses the packets, which is calculated and shown in Fig. 12.

During data transmission, some packets are lost due to the malicious activities and the packet drop ratio of the projected scheme and other techniques are given in Table V and it is represented in Fig. 13.

Here, when considered 50 numbers of packets transmitted then MBDP-AODV attained 23%, OLSR achieves 21% and CLPDM attained 22% Packet drop ratio. Moreover, the proposed ABMZP achieved a low packet drop as 18% validated with other procedures.

5) Throughput calculation: It denotes the rate of transmitting data packets through the network that is distributed through convinced physical or logical links. This value is denoted in bit/s or bps which is represented in Fig. 14.

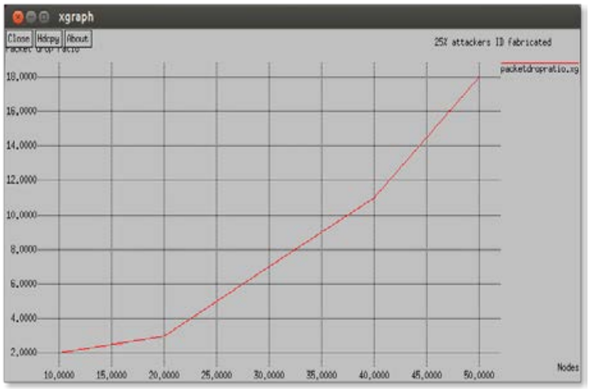


Fig. 12. Calculation of Packet Drop Ratio.

TABLE V. EVALUATION OF PACKET DROP RATIO

	Packet Drop ratio (%)			
No. of Nodes	**MBDP AODV[19]**	**OLSR [29]**	**CLPDM-SI [30]**	**Proposed [ABMZP]**
10	3.5	4	3	2
20	6	4.5	7	3
30	9.5	8	13	7
40	16	13.5	18	11
50	23	21	22	18

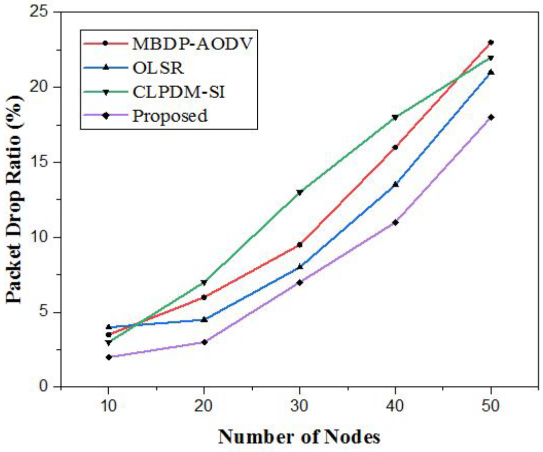


Fig. 13. Comparison of Packet Drop Ratio.

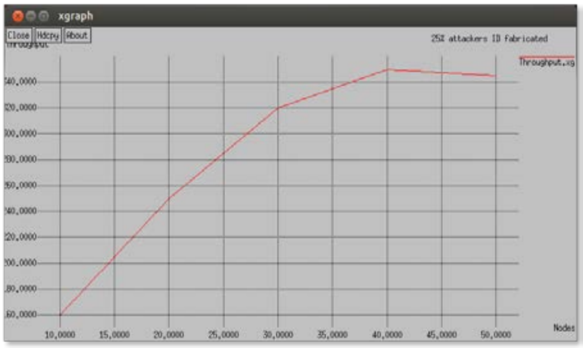


Fig. 14. Calculation of throughput.

The proposed manner achieves high throughput value validated with other prevailing approaches that are detailed in Table VI.

TABLE VI. EVALUATION OF THROUGHPUT

Throughput (kbps)				
Nodes	**MBDP AODV[19]**	**OLSR [29]**	**CLPDM-SI [30]**	**Proposed [ABMZP]**
10	20	36	145	160
20	18	48	198	250
30	17	54	201	320
40	15.5	88	198	350
50	18.63	95	201	345

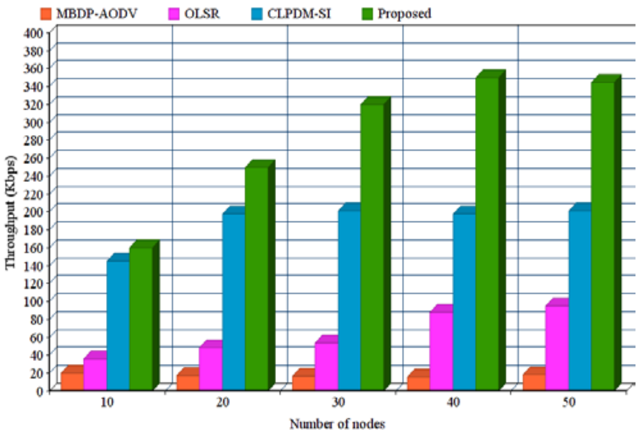


Fig. 15. Comparison of throughput.

Here, some recent methods like MBDP-AODV attained 18.63kbps, OLSR attained 95kbps, and the CLPDM-SI approach accomplished 201kbps for transmitting 50 numbers of nodes. But, the proposed ABMZP achieved 345 kbps high throughput value compared with other techniques, which is represented in the Fig. 15.

D. Discussion

The proposed ABMZP model provides secure communication in the MANET that is transferring the message from the sender hub to the target hub without any interruption. Therefore, the introduced novel ABMZP method provides better outcomes in terms of attack prevention rate, PDR, packet drop ratio, throughput, and delay validated with existing approaches like MBDP-AODV, OLSR, and CLPDM-SI.

VI. Conclusion

Generally, MANET is affected by various attacks such as wormhole attack, BH attack, and GH attack. To protect the information during data transmission the attack prevention is necessary. Hence, this paper introduced the novel prevention method as African Buffalo Monitoring Zone Protocol (ABMZP) for securing the communication. Thus, the ABMZP approach prevents the data from wormhole attacks and other malicious activities in the communication network. Consequently, if ABMZP detects the malicious activities then it alerts the source node. Also, it neglects the attack and provides secure transmission through an optimal path. Hence, it achieves a 99.96% high attack prevention ratio, 99.98% PDR, lower delay, and high throughput ratio.

References

[1] Das, Santosh Kumar, et al. Design Frameworks for Wireless Networks. Springer, 2020.

[2] Gowtham, M. S., and KamalrajSubramaniam. "Congestion control and packet recovery for cross layer approach in MANET." Cluster Computing 22.5 (2019): 12029-12036.

[3] Nehra, Deepa, Kanwalvir Singh Dhindsa, and Bharat Bhushan. "A Security Model to Make Communication Secure in Cluster-Based MANETs." Data Engineering and Communication Technology. Springer, Singapore, 2020. 183-193.

[4] Peng, Wei, and Xicheng Lu. "AHBP: An efficient broadcast protocol for mobile ad hoc networks." Journal of computer science and technology 16.2 (2001): 114-125.

[5] Erdelj, Milan, MichałKról, and Enrico Natalizio. "Wireless sensor networks and multi-UAV systems for natural disaster management." Computer Networks 124 (2017): 72-86.

[6] Yadav, Ajay Kumar, and SachinTripathi. "QMRPRNS: Design of QoS multicast routing protocol using reliable node selection scheme for MANETs." Peer-to-Peer Networking and Applications 10.4 (2017): 897-909.

[7] Lou, Wei, and Jie Wu. "A cluster-based backbone infrastructure for broadcasting in manets." Proceedings International Parallel and Distributed Processing Symposium. IEEE, 2003.

[8] Gautam, Divya, and VrindaTokekar. "A Comparative Study of DoS Attack Detection and Mitigation Techniques in MANET." Social Networking and Computational Intelligence. Springer, Singapore, 2020. 615-626.

[9] Vanitha, K., and AMJ ZubairRahaman. "Preventing malicious packet dropping nodes in MANET using IFHM based SAODV routing protocol." Cluster Computing 22.6 (2019): 13453-13461.

[10] Jamal, Tauseef, and Shariq Aziz Butt. "Malicious node analysis in MANETS." International Journal of Information Technology 11.4 (2019): 859-867.

[11] Bisen, Dhananjay, and Sanjeev Sharma. "Fuzzy based detection of malicious activity for security assessment of MANET." National Academy Science Letters 41.1 (2018): 23-28.

[12] Ghathwan, Khalil I., and Abdul Razak B. Yaakub. "An Artificial Intelligence Technique for Prevent Black Hole Attacks in MANET." Recent Advances on Soft Computing and Data Mining. Springer, Cham, 2014. 121-131.

[13] Rana, Anuj, VinayRana, and Sandeep Gupta. "EMAODV: Technique to prevent collaborative attacks in MANETs." Procedia Computer Science 70 (2015): 137-145.

[14] Jamal, Tauseef, and Shariq Aziz Butt. "Malicious node analysis in MANETS." International Journal of Information Technology 11.4 (2019): 859-867.

[15] Hussain, Mohammed Ali, and D. Balaganesh. "Prevention of Packet Drop by System Fault in MANET Due to Buffer Overflow." Intelligent Computing and Innovation on Data Science. Springer, Singapore, 2020. 615-620.

[16] Gupta, Prakhar, et al. "Reliability factor based AODV protocol: Prevention of black hole attack in MANET." Smart Innovations in Communication and Computational Sciences. Springer, Singapore, 2019. 271-279.

[17] Su, Ming-Yang. "Prevention of selective black hole attacks on mobile ad hoc networks through intrusion detection systems." Computer Communications 34.1 (2011): 107-117.

[18] Vo, Tu T., Ngoc T. Luong, and Doan Hoang. "MLAMAN: a novel multi-level authentication model and protocol for preventing wormhole attack in mobile ad hoc network." Wireless Networks 25.7 (2019): 4115-4132.

[19] Gurung, Shashi, and Siddhartha Chauhan. "A dynamic threshold based algorithm for improving security and performance of AODV under black-hole attack in MANET." Wireless Networks 25.4 (2019): 1685-1695.

[20] Doss, Srinath, et al. "APD-JFAD: accurate prevention and detection of jelly fish attack in MANET." Ieee Access 6 (2018): 56954-56965.

[21] Gayathri, S., et al. "Wormhole Attack Detection using Energy Model in MANETs." 2019 2nd International Conference on Power and Embedded Drive Control (ICPEDC). IEEE, 2019.

[22] Gurung, Shashi, and Siddhartha Chauhan. "A novel approach for mitigating gray hole attack in MANET." Wireless Networks 24.2 (2018): 565-579.

[23] Gautam, Divya, and VrindaTokekar. "Pattern Based Detection and Mitigation of DoS Attacks in MANET Using SVM-PSO." International Conference on Sustainable and Innovative Solutions for Current Challenges in Engineering & Technology. Springer, Cham, 2019.

[24] Gurung, Shashi, and Siddhartha Chauhan. "Performance analysis of black-hole attack mitigation protocols under gray-hole attacks in MANET." Wireless Networks 25.3 (2019): 975-988.

[25] Rajendran, N., P. K. Jawahar, and R. Priyadarshini. "Makespan of routing and security in Cross Centric Intrusion Detection System (CCIDS) over black hole attacks and rushing attacks in MANET." International Journal of Intelligent Unmanned Systems (2019).

[26] Chang, Jian-Ming, et al. "Defending against collaborative attacks by malicious nodes in MANETs: A cooperative bait detection approach." IEEE systems journal 9.1 (2014): 65-75.

[27] Odili, Julius Beneoluchi, MohdNizamMohmadKahar, and Shahid Anwar. "African buffalo optimization: A swarm-intelligence technique." Procedia Computer Science 76 (2015): 443-448.

[28] Selvi, P. Tamil, and C. Suresh GhanaDhas. "A novel algorithm for enhancement of energy efficient zone based routing protocol for MANET." Mobile Networks and Applications 24.2 (2019): 307-317.

[29] Bhuvaneswari, R., and R. Ramachandran. "Denial of service attack solution in OLSR based manet by varying number of fictitious nodes." Cluster Computing 22.5 (2019): 12689-12699.

[30] Bhande, Premala, and M. D. Bakhar. "Cross layer packet drop attack detection in MANET using swarm intelligence." International Journal of Information Technology (2019): 1-10.

[31] Chaitanya, G.K.,Amarendra, K.,Aslam S.,Soundharya, U.L.,Saikushwanth V., “Prevention of data theft attacks in infrastructure as a service cloud through trusted computing” International Journal of Innovative Technology and Exploring Engineering, 2019, 8(6 Special Issue 4), pp. 1278-1283.

[32] Gogineni Krishna Chaitanya and Krovi Raja Sekhar, “GAIT based Behavioral Authentication using Hybrid Swarm based Feed Forward Neural Network” International Journal of Advanced Computer Science and Applications (IJACSA), 11(9), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0110939.

[33] D Toradmalle, J Muthukuru, B Sathyanarayana “Cryptanalysis of an Improved ECDSA”, International Journal of Engineering Research and Technology volume-11,issue-2,pg no:615-619.

[34] Gogineni Krishna Chaitanya and Krovi.Raja Sekhar(in press), “Knowledge-Based Gait Behavioural Authentication Through A Machine Learning Approach” International Journal of Biomedical Engineering and Technology(in press).

[35] Priyadarsini P., Sai M.S.S., Suneetha A., Santhi M.V.B.T.”Robust feature selection technique for Intrusion Detection System”. International Journal of Control and Automation 2018.

[36] Gummadi A, Rao K.R.”EECLA: Clustering and localization techniques to improve energy efficient routing in wireless sensor networks”, Indonesian Journal of Electrical Engineering and Computer Science 2018.

[37] B.Suresh Babu.”Adaptive and Efficient Routing Model for MANET using TSCH Network” Jour of Adv Research in Dynamical & Control Systems” 2018.

[38] Kavitha M., Anvesh K., Arun Kumar P., Sravani P.” IoT based home intrusion detection system”, International Journal of Recent Technology and Engineering 2019.

[39] Sai Harika T., Madhusri N., Varaprasad P.V.V.”Detection, prevention and mitigation of black hole attack for MANET”.International Journal of Recent Technology and Engineering 2019.

[40] Swetha K., Sowmya V., Srihitha K., Adithya D,” A novel technique for secure routing in wireless sensor networks”, International Journal of Innovative Technology and Exploring Engineering 2019.

[41] Bhandari R.R., Rajasekhar K.” Energy-efficient routing-based clustering approaches and sleep scheduling algorithm for network lifetime maximization in sensor network: A survey”, Lecture Notes in Networks and Systems, 2020.

[42] Ananthakumaran. S, Sathishkumar. M, Bhavani. R, Ravinder Reddy. R, “Prevention Of Routing Attacks Using Trust-Based Multipath Protocol,” International Journal Of Advanced Trends In Computer Science And Engineering, Vol. 9, No. 3, Pp. 4022-4029, May-June, 2020.

[43] Gogineni Krishna Chaitanya and Krovi.Raja Sekhar, “A Human Gait Recognition Against Information Theft in Smartphone using Residual Convolutional Neural Network” International Journal of Advanced Computer Science and Applications(IJACSA), 11(5), 2020.http://dx.doi.org/10.14569/IJACSA.2020.0110544.

[44] V Kavidha, S Ananthakumaran -,”Novel energy-efficient secure routing protocol for wireless sensor networks with Mobile sink” Peer-to-Peer networking and Applications 2019.

[45] D Toradmalle, J Muthukuru, B Sathyanarayana,” Certificateless and provably-secure digital signature scheme based on elliptic curve.”- International Journal of Electrical & Computer Engineering (2088-8708)- 2019.

[46] Ananthakumaran.S, Debrup Banerjee, P. G. Om Prakash, R. Bhavani, “Fuzzified Energy Efficient Mechanism (FEEM) in Wireless Sensor Network,” International Journal of Emerging Trends in Engineering Research, Vol. 8, No. 9, pp. 6889-6396, September, 2020.

Analysis of Indonesian Motorcycle Gang with Social Network Approach

Edi Surya Negara*[1], Ria Andryani[2], Deni Erlansyah[3], Rezki Syaputra[4]
Data Science Interdisciplinary Research Center, Faculty of Computer Science
Universitas Bina Darma, Palembang, Indonesia

***Abstract*—Analysis of motorcycle gang networks in Indonesia was conducted to determine the dynamics of the motor gang network. This analysis is needed by the government in making appropriate and effective policies in overcoming social problems caused by the existence of this group. The purpose of this study is to detect and determine the community structure of motorcycle gang networks in Indonesia through the use of big data available on the internet, especially social media. This research also utilizes several approaches such as social and behavioral sciences, as well as the computer technology in understanding and finding solutions to problems that arise in society. This study uses a social network analysis method as an instrument that will reveal the social structure of motorcycle gangs with a network centrality approach and community detection. This research succeeded in finding the network structure pattern and network insight of motorcycle gangs by finding the most influential actors. The study also found 25 motorcycle gang groups with high-value network interactions and these groups had more than 2000 active members on social media. In the biker gang social network analysis, the most influential actor has 531 degrees with a weighted degree of 1557.**

Keywords—Social network; data mining; data analytics; community detection; motorcycle gang

I. Introduction

As of December 2018, the Republic of Indonesia Police Traffic Corps data recorded that 120.1 million units of motorbikes were circulating in the community. Some motorcycle users build motorcycle groups or clubs. The majority of these groups are law-abiding members of the community and are not involved in illegal activities, but a small number are involved in activities that disturb public order and even lead to criminal acts that are carried out in groups or organized. This group of motorcycle riders who have a tendency to carry out activities that disturb public order and break the law is known in Indonesian society by the name of the motorcycle gang.

The criminal case of motorcycle gang members as a social phenomenon requires an in-depth social study as the basis for determining government policy for the prevention and control of crime. Experience in several countries such as the United States, Canada, Australia, and New Zealand shows that groups of motorcyclists who initially only disturbed public order easily turned into Outlaws Motorcycle Gang which is a criminal organization involved in the business of prostitution, human trafficking, arms smuggling, narcotics and drugs, and other serious crimes [1], [2], [3]. The fact that Indonesia is in a narcotics emergency, is not impossible if the motorcycle gang will interact and collaborate with other criminal organizations. To prevent the motorcycle gang which was originally a youth delinquency group transformed into a criminal organization, appropriate and effective policies and responses are needed to anticipate this threat. Furthermore, appropriate and effective policies and responses require an understanding of the behavior, structure of social networks [4], [5], and group dynamics of the motorcycle gang [6].

The increasing number of criminal problems caused by crimes committed by motorcycle gangs is of concern to the government. So that a special approach is needed to be able to provide solutions to these problems. The main problem of this research is how to find out the structure of the social network of motorcycle gang groups in Indonesia. In addition, this research can also provide an overview of how the growth patterns of motorcycle gang groups in Indonesia.

To measure the structure of social networks and group dynamics, a social network analysis approach can be used. In general, a network is a relationship of two or more objects where the object is denoted by a node and the relation between objects is denoted by an edge. The representation of objects on the network can be in the form of people called social networks and the representation of relations can be in the form of social relations such as friendship [20]. Social networks are social relationships between individuals or groups that interact with each other. This relationship can be a relationship between individuals and individuals, individuals with groups, or groups with groups [7], [8]. Social Network Analysis is a study that is used to look at human social relations through graph theory [9]. Through graph theory, social network analysis can find out the structure of social relations in certain groups through degree centrality and community detection. The main task of social networking analysis is to identify the most influential actors in a social network using statistical measures [10].

Degree Centrality is a measure in the graph used in network analysis to find important structures of vertices and edges. Centrality generally determines the importance of a node-based only on the structure of the graph. The simplest definition of node centrality is that the central node must be the most active node or the node that has the most ties with other nodes in the network [11]. The formula used to calculate the degree of centrality of a node in a graph can be done using the formula in equation (1). Where *d(ni)* is the number of relations that a node has with another node in the graph.

$$CD(ni) = d(ni) \quad (1)$$

*Corresponding Author

Community detection is one method of social network analysis in detecting communities on a network. Implicitly the definition of community refers to the structure of the network is a group of vertices that are connected to one another with certain proximity and density that have similarities in the network. A community, also known as a cluster, is usually considered a group of nodes that have many connections to each other and multiple connections to the entire network. Identifying communities in a network can provide valuable information about the structural properties of the network, interactions between nodes in the community, and the role of nodes in each community [6]. To detect communities on social networks, the formula modularity can be used. Modularity shows how relationships that exist on social networks can form different groups in a social network. The modularity formula can be seen in equation (2). Where Q is modularity, e_{ii} is probability edge is in module i, and a_i^2 is probability a random edge would fall into module.

$$\mathrm{Q} = \sum_{1=1}^{k}(e_{ii} - a_i^2) \tag{2}$$

Improved social network analysis methods are characterized by the ability to identify. Using techniques such as block modeling can reveal interactions between groups and between criminal networks by determining the presence or absence of associative bonds, identifying networks involved in one or more criminal activities, and then identifying relationships between groups or individuals (chain structure) that will be seen through clustering [15]. Subgroups, leaders, group members, main leaders are identified through degrees, interactions, and closeness [12]. Social networking analysis is the main tool used in studying and investigating criminal organizations and radical groups [13], [14].

In this study, we used a social network analysis approach using data from the social networking site Facebook. Facebook data is used because currently Facebook is the social media platform with the largest number of users compared to other social media platforms. There are four important reasons that use the social network analysis approach to study the dynamics and social networks of motorcycle gang groups in Indonesia. First, based on the nature and social behavior of the motorcycle gang analysis social networking is a very suitable method for studying the behavior and dynamics of this group. Second, the demographics of motorcycle gang members who average age 15-18 years [16] are included in the age of active users of social media. The demographics of motorcycle gang members who are a subset of active social media user demographics enable the use of social network analysis techniques, not only as data collection techniques for analyzing motorcycle gang network phenomena but also to analyze, monitor, and even stimulate the dynamics of motorcycle gang groups. Third, the number of users of social networks in Indonesia. Indonesia is currently the country with the fourth-largest number of Facebook users in the world, with 60.3 million users in May 2014. The chances of a motorcycle gang member who is also an active Facebook user are huge. Fourth, modern computers that have high-performance memory and processor. Although social network analysis was first developed in the 1930s it can only be applied to analyze social problems involving real networks with a large number of actors about two decades ago because there are no computers that have memory and processors that can handle the complexity of the problem that must be resolved. A large number of motorcycle gang members using social media provides an opportunity to see patterns, social network structures that occur among group members by utilizing social networking analytical methods.

In contrast to research on social problems that use the science of sociology, criminology, or conventional psychology that relies on data collection using surveys or ethnography to uncover the effects and causes of these problems, this study uses the social networking paradigm. Analysis of social networks provides an alternative view where the attributes possessed by individuals are reduced compared to relationships and bonds with other actors who are in the network. Analysis of social networks has been used to explain social phenomena and organizational behavior see [16],[17] and bibliography therein.

In this article, it is explained about the social network structure of motorcycle gangs in Indonesia, where the data used in this study is the data of motorcycle gangs from several cities in Indonesia, namely: Jakarta, Bandung, and Palembang. The reason for choosing motorcycle gangs from the cities of Jakarta, Bandung, and Palembang as a dataset to be analyzed is due to simple content analysis on Indonesian-language online news sites about motorcycle gang activity from 2012 - 2020 the three cities are often recorded on online news sites related to actions criminal motorcycle gang [18]. The analysis of social networks will provide insights into central actors with the degree centrality method and community detection with the modularity method [19], [6].

II. Research Method

Analysis of Indonesia Motorcycle Gang with Social Network Approach. This study consists of four iterative stages, namely Data Collection, Data Processing, Network Analysis, and Graph Visualization. The flow of the research process can be seen in Fig. 1.

A. Data Collection

This research uses data sourced from Facebook. The data taken is Facebook Fanpage data from the motorcycle community in the cities of Jakarta, Bandung, and Palembang. Data retrieved with web crawling techniques using the Netvizz application. Netvizz is an application that functions to extract data from the Facebook platform. Data attributes in this study are Member_ID, Label, Type_Post, and Comment. To build a motorcycle gang community data network, the "Member_ID" attribute will be denoted as a node, then the "status" attribute will be denoted as the edge. Data representation of motorcycle gang network groups can be seen in Fig. 2. Data generated from this process can be seen in Table I.

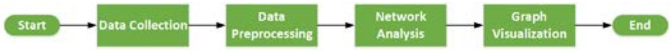


Fig. 1. Process Flow Research Stages.

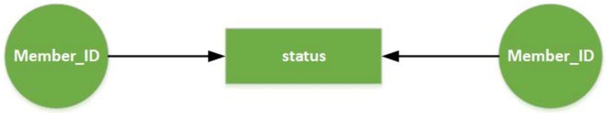


Fig. 2. Network representation of Indonesian Motorcycle Gang Community.

TABLE I. NUMBER OF DATA

Data Source	Number of Node	Number of Edge
Jakarta	2126	5730
Bandung	2138	4087
Makasar	1297	4295

B. Data Preprocessing

At this stage the data will be cleared by removing rows of no value, transforming the Member_ID, label, Type_Post, and Comment attribute to numeric, removing the leading and ending spaces, deleting all punctuation, changing all letters into lowercase letters, and repeating "FOR" To add a new line according to the number of" Comments "owned by the node. Fig. 3 shows some of the data that has been cleared at the data preprocessing stage.

C. Network Analysis

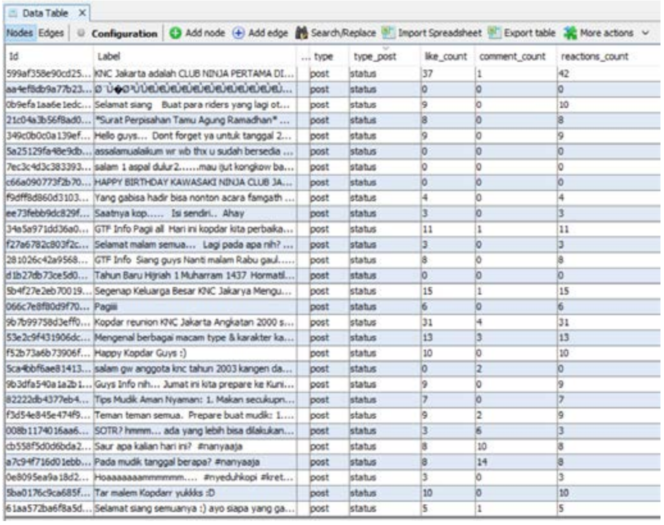
The formulation and formation of the motorcycle gang network is the first step that must be done before conducting network analysis. In social networks, there are two forms of relationships that occur namely: interconnection and interaction. Interconnection networks are formed by a node connection that is part of the same group in a network. Network interaction is formed from node relationships that occur based on communication interactions between nodes in the network through chat and respot to posts.

After the network is formed, the next step is to conduct a motorcycle gang network analysis. To conduct an analysis, this study uses the degree centrality, betweenness centrality, closeness centrality and community detection algorithms on social networks to find out the central actors or actors who have a big influence on the motorcycle gang, and recognize the motorcycle gang community in Indonesia based on recording interconnection and interaction happened to social media.

To analyze the social network of motorcycle gangs in Indonesia, iteratively, the framework or stages of analysis are built on the basis of the centrality and community detection analysis approach. In detail, these steps are shown in Fig. 4.

D. Graph Visualization

At this stage, the data that has gone through the preprocessing and processing stages will then be visualized using graphs, with the size discussed in the previous chapter, namely degree centrality, betweenness centrality, closeness centrality, and community detection. At this stage, you will also find the most influential node in the graph. Visualization is done using Gephi Graph Visualization 0.9.2 software.



Id	Label	...	type	type_post	like_count	comment_count	reactions_count
599af358e90cd25...	KNC Jakarta adalah CLUB NINJA PERTAMA DI...		post	status	37	1	42
aa4ef8db9a77b23...	Ø Ù�Ø³ÙÙ€ÙÙ€ÙÙ€ÙÙ€ÙÙ€ÙÙ€...		post	status	0	0	0
0b9efa1aa6e1edc...	Selamat siang Buat para riders yang lagi ot...		post	status	9	0	10
21c04a3b56f8ad0...	*Surat Perpisahan Tamu Agung Ramadhan* ...		post	status	8	0	8
349c0b0c0a139ef...	Hello guys... Dont forget ya untuk tanggal 2...		post	status	9	0	9
5a25129fa48e9db...	assalamualaikum wr wb thx u sudah bersedia ...		post	status	0	0	0
7ec3c4d3c383393...	salam 1 aspal dulur2......mau ijut kongkow ba...		post	status	0	0	0
c66a090773f2b70...	HAPPY BIRTHDAY KAWASAKI NINJA CLUB JA...		post	status	0	0	0
f9dff8d860d3103...	Yang gabisa hadir bisa nonton acara famgath ...		post	status	4	0	4
ee73febb9dc829f...	Saatnya kop...... Isi sendiri.. Ahay		post	status	3	0	3
34a5a971dd36a0...	GTF Info Pagi all Hari ini kopdar kita perbaika...		post	status	11	1	11
f27a6782c803f2c...	Selamat malam semua... Lagi pada apa nih? ...		post	status	3	0	3
281026c42a9568...	GTF Info Siang guys Nanti malam Rabu gaul.....		post	status	8	0	8
d1b27db73ce5d0...	Tahun Baru Hijriah 1 Muharram 1437 Hormati...		post	status	0	0	0
5b4f27e2eb70019...	Segenap Keluarga Besar KNC Jakarya Mengu...		post	status	15	1	15
066c7e8f80d9f70...	Pagiii		post	status	6	0	6
9b7b99758d3eff0...	Kopdar reunion KNC Jakarta Angkatan 2000 s...		post	status	31	4	31
53e2c9f431906dc...	Mengenal berbagai macam type & karakter ka...		post	status	13	3	13
f52b73a6b73906f...	Happy Kopdar Guys :)		post	status	10	0	10
5ca4bbf6ae81413...	salam gw anggota knc tahun 2003 kangen da...		post	status	0	2	0
9b3dfa540a1a2b1...	Guys Info nih... Jumat ini kita prepare ke Kuni...		post	status	9	0	9
82222db4377eb4...	Tips Mudik Aman Nyaman: 1. Makan secukupn...		post	status	7	0	7
f3d54e845e474f9...	Teman teman semua. Prepare buat mudik: 1....		post	status	9	2	9
008b1174016aa6...	SOTR? hmmm.... ada yang lebih bisa dilakukan...		post	status	3	6	3
cb558f5d0d6bda2...	Saur apa kalian hari ini? #nanyaaja		post	status	8	10	8
a7c94f716d01ebb...	Pada mudik tanggal berapa? #nanyaaja		post	status	8	14	8
0e8095ea9a18d2...	Hoaaaaaaammmmmm.... #nyeduhkopi #kret...		post	status	3	0	3
5ba0176c9ca685f...	Tar malem Kopdarr yukkks :D		post	status	10	0	10
61aa572ba6f8a5d...	Selamat siang semuanya :) ayo siapa yang ga...		post	status	5	1	5

Fig. 3. Clean Dataset of the Indonesian Motorcycle Gang Community.

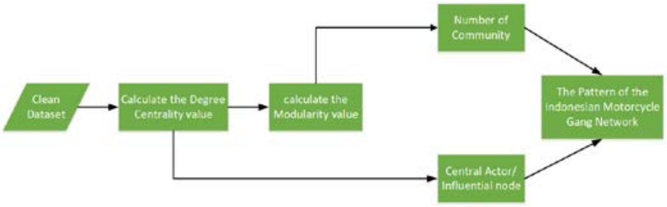

Clean Dataset
Calculate the Degree Centrality value
calculate the Modularity value
Number of Community
Central Actor/ Influential node
The Pattern of the Indonesian Motorcycle Gang Network

Fig. 4. Method for Network Analysis of Indonesia Motorcycle Gang Community.

III. RESULT AND DISCUSSION

Analysis conducted on the motorcycle group network conducted in this study has been able to produce a pattern of motorcycle gang social networks in Indonesia with some insights generated from these patterns. The insights generated from the network analysis are the central actors in the motorcycle gang through the degree centrality algorithm and this analysis also finds groups of motorcycle gangs in Indonesia through interconnection studies and interactions of each member through the community detection algorithm.

A. Degree Centrality

To measure the central actor of a node in a network can use the Degree Centrality. The results of calculating the centrality of the top 5 nodes can be seen in Table II.

Member_ID *afb4c17f761de7e2546b764be5b30fca3132e8b* has the highest centrality value with 531 degree. This can be interpreted that the actor is an actor who is in the motorcycle gang community is quite active in interaction through other user posts. In addition, high centrality values can also be found in the "post" variable. This can be interpreted that posts in the form of "status" submitted by the user get a high response through "comment" replies from other users. Visual graph based on Degree Centrality can be seen in Fig. 5.

TABLE II. CENTRALITY DEGREE VALUE

Member_ID	Type	Type_ Post	Degree	Weighted Degree
fafb4c17f761de7e2546b764be5b30fca3132e8b	User	User	531	1557.0
37ca158055c0c4c39dea128777c2f9a5d81c21de	User	User	299	470.0
920ac5756fd6bce87613879f058203878492caaf	Post	Status	287	293.0
62c1bbb3c4b26ec6ae49dc65b883bbe789de61ca	User	User	276	475.0
1fe624bab07d77e915f27c1ae8babd16ad7eccb7	User	User	189	329.0

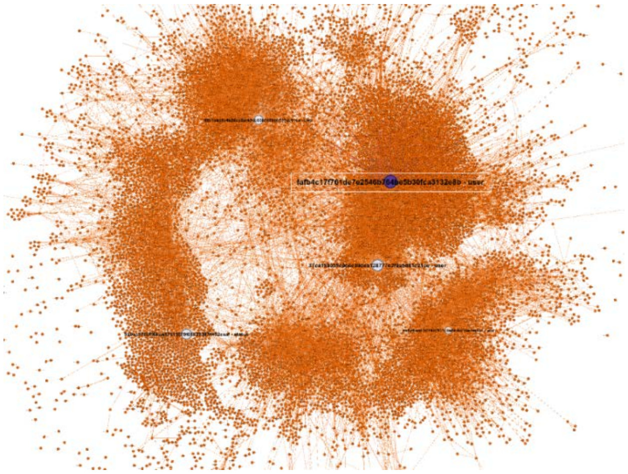


Fig. 5. Graph Visualization based on Degree Centrality.

B. Betweenness Centrality

This measure can be used to measure the influence of nodes in the network. From the popularity level it is known that the most popular node in the network is the node with Member_ID fafb4c17f761de7e2546b764be5b30fca3132e8b. The node with the highest betweenness centrality value because this node is a connecting node or hub between other cluster networks in the network. The most influential node in the motorcycle gang community is the node with Member_ID *37ca158055c0c4c39dea128777c2f9a5d81c21de*. This can be proven by the results of the measurement of Betweenness Centrality in Table III. Node with Member_ID *37ca158055c0c4c39dea128777c2f9a5d81c21de* which is the best link between the other nodes. This node has the most influence on other nodes. Visual results can be seen in Fig. 6.

C. Closeness Centrality

In this study, the graph used is an undirect graph. Then the Closeness calculation does not include InCloseness and OutCloseness. The results of calculating the Closeness Centrality value and Harmonic Closeness of the top five nodes can be seen in Table IV. The closeness centrality measure in this study aims to measure the centrality in the network, where the centrality measure is calculated as the inverse of the total length of the shortest path between the node and all other nodes in the graph. Harmonic centrality is a variant of closeness centrality that was invented to solve the problem the original formula had when dealing with unconnected graphs.

From the table above it can be seen that the highest closeness value is the node with Member_ID *37ca158055c0c4c39dea128777c2f9a5d81c21de*, and the difference in closeness value between nodes in Table IV is not much different. This shows that the node with id has a high closeness value with other nodes, with the interpretation that the node is able to interact quickly with many other nodes. The closeness centrality value can be interpreted as a measure of the closeness between one user and another on the motorcycle gang network based on the average length of one user to all users in the network.

D. Community Detection

Based on its definition that a community is a social group that shares an environment, a group can be interpreted as a collection of individuals. Therefore, to be able to meet these criteria, filtering is needed for nodes that have no relationship with other nodes (standalone). Node filtering uses the degree range in the gephi application to filter nodes that have no relationship, after which modularity calculations are performed to detect the community. Then the modularity calculation results obtained by 0.771 and the number of communities as much as 25 with a minimum degree range 1. The results can be seen in Fig. 7.

TABLE III. BETWEENNESS CENTRALITY VALUE

Member_ID	Type	Type_Post	Betweenness Centrality
37ca158055c0c4c39dea128777c2f9a5d81c21de	User	User	1.1028731076216102E7
fafb4c17f761de7e2546b764be5b30fca3132e8b	User	User	8533454.165279571
62c1bbb3c4b26ec6ae49dc65b883bbe789de61ca	User	User	3565699.073132527
920ac5756fd6bce87613879f058203878492caaf	Post	Status	3113989.7875725445
ffdc0a318547cb46436b568fdde11538f7bb0308	User	User	3103838.9869805044

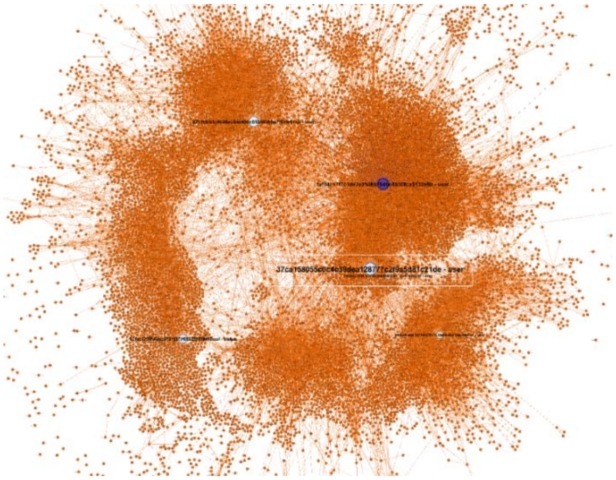


Fig. 6. Graph Visualization based on betweenness Centrality.

TABLE IV. CLOSENESS CENTRALITY VALUE

Member_ID	**Closeness Centrality**	**Harmonic Closeness**
37ca158055c0c4c39dea128777c2f9a5d81c21de	0.293624	0.335476
fafb4c17f761de7e2546b764be5b30fca3132e8b	0.27157	0.326414
189d8fa179d15fc843283744099722ba1db46609	0.268945	0.292587
ba9a4aa1b36216f9560358e38fee246625467b41	0.26634	0.293028
6f0fdfc005a5c10625580b88d6f2a6caec69591b	0.263916	0.289941

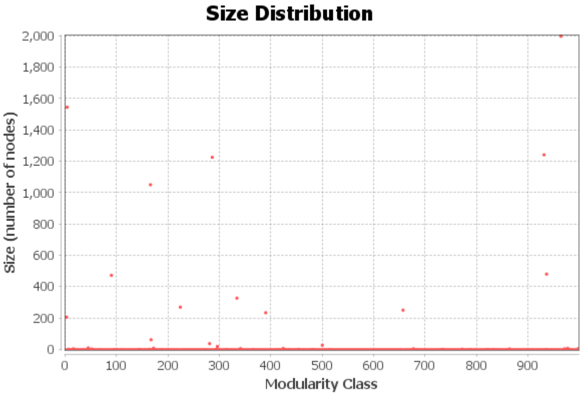


Fig. 7. Modularity Class.

From Fig. 7 it can be seen that the largest community is class 25 with the highest number of nodes or more than 2000 nodes. Fig. 7 is generated from the visualization of the results of community calculations on the motorcycle gang dataset using the modularity algorithm approach. The results of the community with 25 classes are community clusters that have been generated from the calculation of modularity where the number of nodes in the class is 2000 users. The picture shows several communities with different color labels for each community. More color distribution at the node shows the percentage of the number of community members or the number of nodes. The community graph can be seen in Fig. 8.

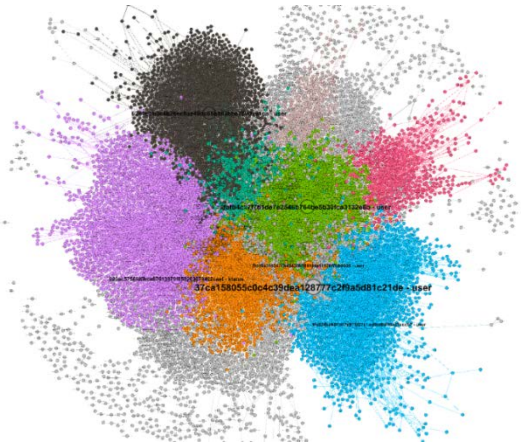


Fig. 8. Visualization of Motorcycle Gang Community Networks.

IV. CONCLUSION

Analysis of the social network of motorcycle gangs in Indonesia has succeeded in providing knowledge about the structure of social networks and providing insights from existing motorcycle gang networks in Indonesia. This research has been able to provide important information for the government in solving the problem of motor gang crime through a social network analysis approach. This analysis finds the most influential actors in the motorcycle gang through high centrality and interaction values. The study also found 25 motorcycle gang groups with high-value network interactions and these groups had more than 2000 active members on social media. In a motorcycle gang, according to the analysis of social networks, the most influential actor has 531 degrees with a weighted degree of 1557. This shows that the actor is the central actor.

Suggestions for the next research are to increase the amount of data to be analyzed using data from various other social media platforms. In addition, another challenge that needs to be resolved is to analyze the motorcycle gang network in real-time using a machine learning approach.

REFERENCES

[1] J. F. Quinn, "Angels, bandidos, outlaws, and pagans: The evolution of organized crime among the big four 1% motorcycle clubs," Deviant Behavior, vol. 22, no. 4, pp. 379–399, 2001.

[2] J. Quinn and D. Shane Koch, "The nature of criminality within one percent motorcycle clubs," Deviant Behavior, vol. 24, no. 3, pp. 281–305, 2003.

[3] D. Shields, "The infamous 'one percenters': A review of the criminality, subculture, and structure of modern biker gangs," Justice Policy Journal, vol. 9, no. 1, pp. 1–33, 2012.

[4] D. N. S. Diah Novita Sari, S. Dedy, and E. S. Negara, "Structure community analysis on social network," in The 6th International Conference on Information Technology and Business Application (ICIBA2017), vol. 1. Penerbit: Pusat Penerbitan dan Percetakan Universitas Bina Darma Press PPP, 2017, pp. 1–7.

[5] A. isnaini Sugiarta, E. S. Negara et al., "Analisis sentralitas aktor pada struktur jaringan politik dengan menggunakan metode social network analysis (sna): Studi kasus group facebook lembaga survei sosial media," in Seminar Nasional Teknologi Informasi dan Komunikasi (SEMNASTIK), vol. 1, no. 1, 2018, pp. 203–209.

[6] E. S. Negara and R. Andryani, "A review on overlapping and non overlapping community detection algorithms for social network analytics," Far East Journal of Electronics and Communications, vol. 18, no. 1, pp. 1–27, 2018.

[7] C. K.-S. Leung, I. J. Medina, and S. K. Tanbeer, "Analyzing social networks to mine important friends," in Social media mining and social network analysis: Emerging research. IGI Global, 2013, pp. 90–104.

[8] M. Ria Andryani, M. Kom, M. Ria Andryani, M. Kom, S. N. Edi et al., "Network of friends to the other friends by social media on facebook," The Turkish Online Journal of Design, Art and Communication, vol. 12, no. 12, pp. 1363–1378, 2017.

[9] M. Tsvetovat and A. Kouznetsov, Social Network Analysis for Startups: Finding connections on the social web. " O'Reilly Media, Inc.", 2011.

[10] M. Cordeiro, R. P. Sarmento, P. Brazdil, and J. Gama, "Evolving networks and social network analysis methods and techniques," Social Media and Journalism: Trends, Connections, Implications, vol. 101, no. 2, 2018.

[11] L. Freeman, "The development of social network analysis," A Study in the Sociology of Science, vol. 1, p. 687, 2004.

[12] A. Sergi, "Case study 4: United kingdom and the activity model," in From Mafia to Organised Crime. Springer, 2017, pp. 177–213.

[13] A. Kriegler, "Using social network analysis to profile organised crime," Policy Brief, vol. 57, 2014.

[14] G. Mastrobuoni and E. Patacchini, “Organized crime networks: An application of network analysis techniques to the american mafia,” Review of Network Economics, vol. 11, no. 3, 2012.

[15] Y. Lu, X. Luo, M. Polgar, and Y. Cao, “Social network analysis of a criminal hacker community,” Journal of Computer Information Systems, vol. 51, no. 2, pp. 31–41, 2010.

[16] I. Tofail, “Tinjauan kriminologis terhadap kejahatan yang dilakukan oleh geng motor di kabupaten gowa,” Makassar: Skripsi Fakultas Hukum Universitas Hasanuddin, 2013.

[17] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, “Network analysis in the social sciences,” science, vol. 323, no. 5916, pp. 892–895, 2009.

[18] Kompas. (2020) Geng-motor. [Online]. Available: https://www.kompas.com/tag/geng-motor.

[19] D. F. Brianna, E. S. Negara, and Y. N. Kunang, “Network centralization analysis approach in the spread of hoax news on social media,” in 2019 International Conference on Electrical Engineering and Computer Science (ICECOS). IEEE, 2019, pp. 303–308.

[20] Negara, E.S., Kerami, D., Wiryana, I.M., Maulana Kusuma, T.B. (2017), “Researchgate data analysis to measure the strength of Indonesian research”. Far East Journal of Electronics and Communications, 17(5), pp.1177-1183.

Predictive System of Semiconductor Failures based on Machine Learning Approach

Yousef El Mourabit[1], Youssef El Habouz[2], Hicham Zougagh[3], Younes Wadiai[4]
TIAD Laboratory, Sciences and Technology Faculty, Sultan Moulay Slimane University, Beni Mellal, Morroco[1, 3, 4]
2Igdr Umr 6290 Cnrs- Rennes1 University, Rennes, France[2]

***Abstract*—Maintenance in manufacturing has been developed and researched in the last few decades at a very rapid rate. It's a major step in process control to build a decision tool that detects defects in equipment or processes as quickly as possible to maintain high process efficiencies. However, the high complexity of machines, and the increase in data available in almost all areas, makes research on improving the accuracy of fault detection via data-mining more and more challenging issue in this field. In our paper we present a new predictive model of semiconductor failures, based on machine learning approach, for predictive maintenance in industry 4.0. The framework of our model includes: Dataset and data acquisition, data preprocessing in three phases (over-sampling, data cleaning, and attribute reduction with principal component analysis (PCA) technique and CfsSubsetEval technique), data modeling, evaluation model and implementation model. We used SECOM dataset to develop four different models based on four algorithms (Naive Bayesian, C4.5 Decision tree, Multilayer perceptron (MLP), Support vector machine), according to the five metrics (True Positive rate, False Positive rate, Precision, F-Mesure and Accuracy). We implemented our new predictive model with 91, 95% of accuracy, as a new efficient predictive model of semiconductor failures.**

Keywords—Machine learning; semiconductor; predictive maintenance; industry 4.0

I. Introduction

Currently, the industrial competition, the huge demand and the digital transformation have encouraged most industries to exploit and take advantage of the available technological tools. Many researches have proved the potential of the Artificial Intelligence (AI) for more efficiency and quality, reducing cost, and improving predictive maintenance services in manufacturing [1]. Nowadays, the industry is gradually developing towards what experts have called Industry 4.0, (The Fourth Industrial Revolution). This fact is strongly associated with the integration between digital and physical systems of production environments. This integration allows the collection of a huge amount of data by different equipment, located in many sectors of the factories [2]. Industry 4.0 is about performing tasks right on time, simultaneously, more efficiently, with more flexibility in a safer and respectful way. Likewise the industry 4.0 technologies integrate machines, products and people, allowing faster and more secure exchange of information [3]. The introduction of new technologies and new services associated with Industry 4.0 revolutionizes many industrial applications, and approaches, such as those in factories regarding automation and predictive industrial maintenance to create smarter work environments, in order to found opportunities for newcomers to de-liver innovative solutions that change business models.

Every year, a huge amount of data is collected by industrial systems, it contains precious information about processes and breakdown that occur in the production. In addition, analyzing and processing these data can show up valuable information and knowledge from system dynamics and manufacturing process [4]. Using various approaches based on data, it is possible to find illustrative results for strategic decision-making, providing advantages such as, increased production, machine fault reduction, maintenance cost reduction, among others [5] [6] [7]. The advantages above have strong relation with maintenance procedures. In manufacturing, equipment maintenance is a very important key, it affects the efficiency and operation time of equipment. As well, equipment faults need to be identified and solved, without production processes shutdown [8].

In literature, various groups and categories of maintenance management strategies can be found. Based on [9] [10], the maintenance procedures are classified as follows: Run-to-Failure (R2F) or Corrective maintenance, Preventive Maintenance (PvM) Time-based maintenance or Scheduled maintenance and Predictive Maintenance (PdM). PdM uses predictive tools to identify when maintenance actions are necessary. Therefore, it permit the early detection of failures by predictive tools using collected data with engineering approaches, statistical inference methods and machine learning techniques. To contribute to this challenge, in this paper we present a new predictive model of semiconductor failures, based on machine learning approach.

The aim of our work is to create a powerful predictive model of semiconductor failures that can predict future events to avoid failures. We used the SECOM dataset, after the preprocessing phases, we compare four predictive models based respectively on Naive Bayesian (NB), C4.5 decision tree, multilayer perceptron (MLP) , support vector machine (SVM) algorithms, in order to implement the most efficient and accurate model. According to several metrics (True Positive rate, false Positive rate, Precision, F-Measure and accuracy) we implemented a new efficient model based on MLP algorithm for predicting equipment failures during the wafer manufacturing process in the semiconductor industry, reached 91% of accuracy. The remainder of the paper is organized as follows: Section 2 introduces the related work. Section 3 presents our approach with detailed framework.

Section 4 details the experiment with a discussion of results. Section 6 concludes with future research directions.

II. Related Work

Artificial Intelligence transforms the traditional factory into a digital paradigm by increasing technological tools, real-time connectivity, and analytics capabilities. Moreover, data become a source of value to find illustrative results for strategic decision-making, in order to identify when maintenance actions are necessary, Hashemian and Bean [11] confirms that the few researches in the PdM area are due to the difficulty and complexity of implementing efficient PdM strategies in production environments. Also, the lack of use of Machine learning (ML) algorithms in PdM applications is related to availability of historical data in equipment failures, and especially, having professionals in the data science and ML field on the production line.

According to [12], Random forest (RF) is a supervised learning algorithm for regression tasks and classification. RF have shown more efficiency when the number of variables is larger than the number of samples. The main contributions of the Canizo work [13], are automation and scalability, also speed in data processing. Its results show an improvement of 5.54%, according to predictive accuracy, when compared to the Kusiak & Verma work [14]. The research developed by Su and Huang [15], presents a predictive fault detection system “HDPass”, in order to perform hard disk drive faults. Using RF algorithm, the result presented by this work is promising, since it achieves 85% of accuracy. Authors used a type of SVM for regression purposes in [16]: Support Regression Vector (SVR). In this work, a modified regression kernel is presented to prognostic problems. In spite of that, the work does not perform any comparison between other ML methods. Results show that the proposed SVR model outperforms a standard SVR model.

Artificial intelligence, within, ML become a powerful tool for developing efficient predictive algorithms in various applications. ML approaches have the ability to deal with multivariate and high dimensional data in dynamic and complex environments [17]. Thus, ML offers powerful approaches for PdM applications. However, the efficiency of these applications depends on the adequate choice of the ML approach. Therefore, the aim of this paper is to present a new predictive model of semiconductor failures, based on the comparison of the most efficient machine learning algorithms (most used in PdM) according to various metrics.

III. Proposed Approach

Data preparation is the first critical phase in the development of a predictive model, it’s an essential step aims converting various types and forms of data into an appropriate format, which is relevant to the predictive model based on machine learning. On the semiconductor manufacturing process, a huge amount of data is collected regularly during processing.

An experimental implementation was conducted to verify the efficiency and performance of the proposed failure prediction model by using SECOM dataset [18]. This dataset consists of 1567 data record and 591 attributes; it’s collected from a semiconductor manufacturing process by monitoring the sensors and the process measurement point. Each record is a vector of 590 sensor measurements in addition to the data of the remaining feature were represented by Pass and Fail (label). Fig. 1 shows the proposed approach for generating a predictive model.

A. Oversampling Phase

The unbalanced distribution of data is a big challenge for standard learning algorithms. In SECOM dataset the number of successful tests is very important (1463 in-stances), compared to the number of failure tests, which is very infrequent (104 instances), this imbalance failure and success record, also the huge number of metrology data obtained from various sensors makes this dataset difficult to evaluate accurately. Therefore the forecasting model needs a data sampling method that can solve the imbalance of the records, for this we propose the sampling method [19].

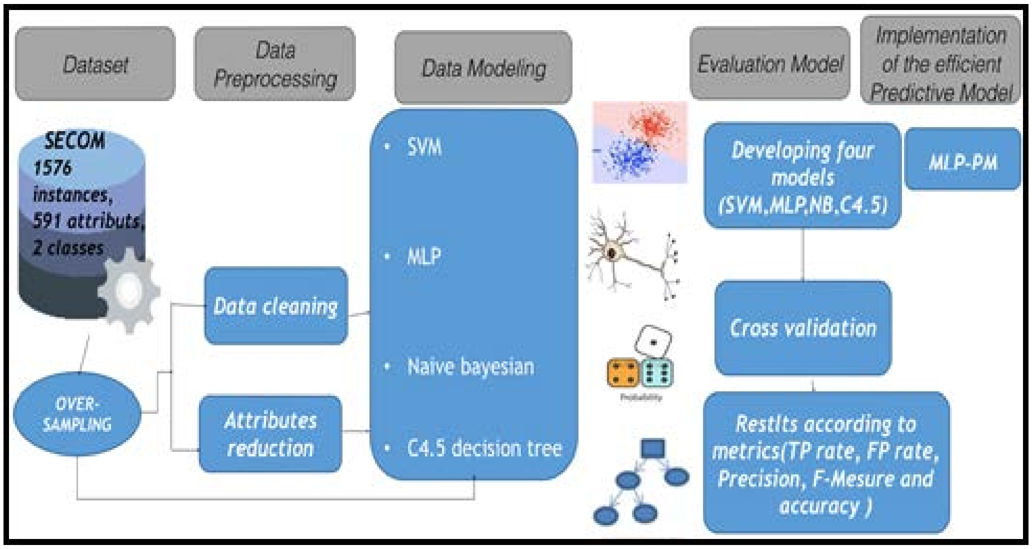

Dataset
Data Preprocessing
Data Modeling
Evaluation Model
Implementation of the efficient Predictive Model
SECOM 1576 instances, 591 attributs, 2 classes
OVER-SAMPLING
Data cleaning
Attributes reduction
• SVM
• MLP
• Naive bayesian
• C4.5 decision tree
Probability
Developing four models (SVM,MLP,NB,C4.5)
MLP-PM
Cross validation
Restlts according to metrics(TP rate, FP rate, Precision, F-Mesure and accuracy)

Fig. 1. Approach Architecture.

B. Data Cleaning Phase

Firstly we check the found value of each attribute, if the data seems to be unique value, i.e. the same value for all records, we remove this feature. Secondly we count in each column the missing data; if it reaches more than 55% we remove this attribute. We removed 158 attributes, and only kept 434 attributes.

C. Attributes Selection / Reduction Phase

A Huge amount of datasets are increasingly widespread in various disciplines. Characteristic selection or dimensionality reduction techniques are required to perform such datasets, also to improve prediction and computation performance, while preserving most information in the data. In this phase we used two methods to compare the results and implement the efficient one. Firstly we apply the CfsSubsetEval (Correlation-based Feature Selection) selection method [20], with the Best First search strategy that evaluates value of a subset of attributes according to the individual predictive capacity of each characteristic and the degree of redundancy.

The subsets of characteristics strongly correlated with the class while having a low intercorrelation are preferred. The result shows that just 17 attributes are considered, adding the label Pass / Fail, we obtain 19 attributes, we save the result as a separate dataset.

Secondly, we used principal component analysis (PCA) [21]. It aims to reduce the dimensionality of a dataset, and preserve as much as possible statistical information and variability. PCA geometrically projecting data onto lower dimensions named principal components (PCs), in order of finding the better summary of the data using a few number of PCs. Fig. 2 shows the correlation matrix of the used dataset.

After the standard scalar normalization, in order to normalize our set of features, we selected 168 best features to maintain approximately 95% of the accumulated variances as shown in Fig. 3.

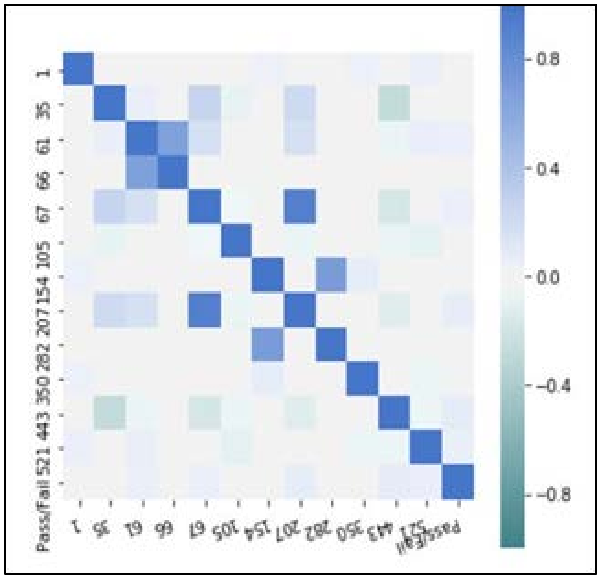


Fig. 2. Correlation Matrix of the used Dataset.

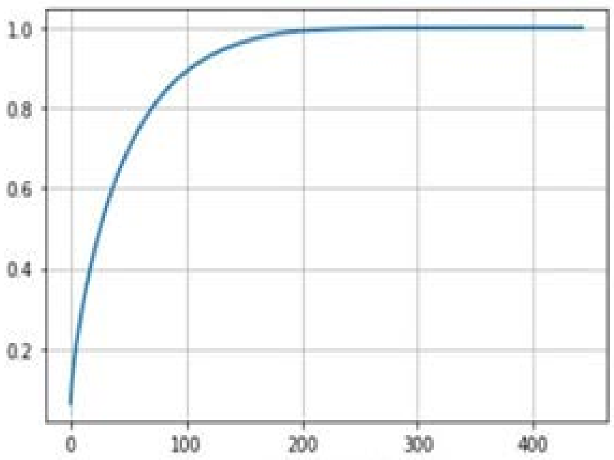


Fig. 3. Commutative Variance.

IV. Experiment and Results

After the preprocessing part, we performed a series of experiments in order to obtain the most efficient predictive model. Firstly, we used four different sets according to the data preprocessing phase: uncleaned dataset, cleaned dataset, cleaned dataset with attributes selected by the CfssubsetEval method, and cleaned attribute reduced by the PCA method.

Secondly, for every dataset, we applied four machine learning algorithms (SVM, NB, MLP, C4.5) [22], and perform the efficiency of these models based on five relevant metrics (TP Rate, FP rate, F-Measure, Precision, and accuracy) [23] [24]. Finally, we implement the most performant and efficient predictive model, based on MLP method using python environment. According to the results below, we can visualize the performance evaluation between the four different machine learning models, using four different dataset.

According to the results above (Fig. 4 to Fig. 8), it’s clear that the using of a dataset with features reduction methods, improve significantly the accuracy of the four predictive models. Moreover, in this case, PCA method shows considerable performance compared to the CfssubsetEval method.

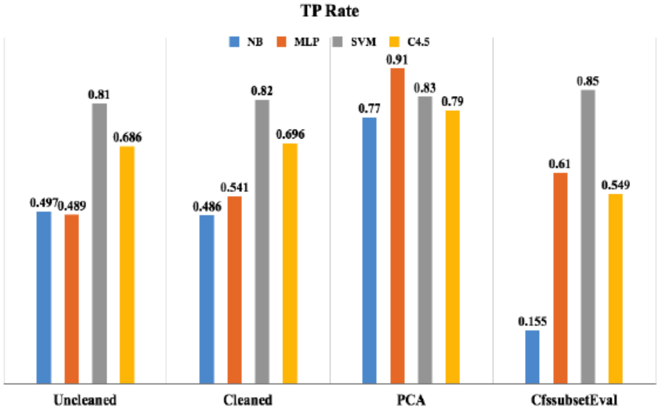


Fig. 4. TP Rate of Machine Learning Models according to the Four Datasets.

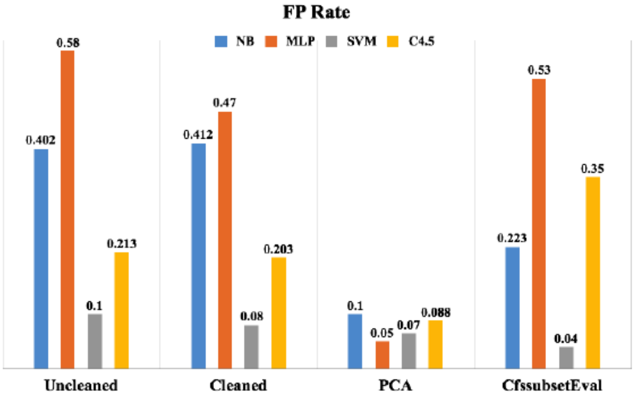


Fig. 5. FP Rate of Machine Learning Models according to the Four Datasets.

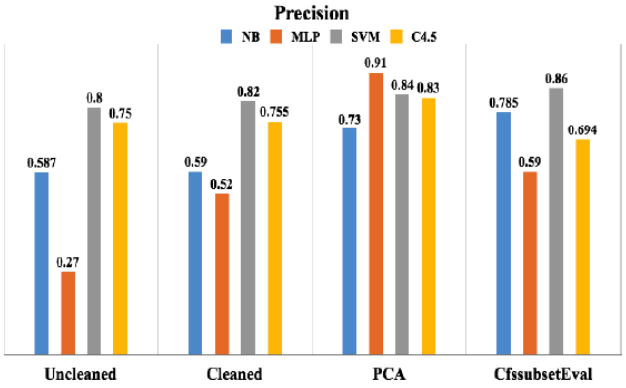


Fig. 6. Precision of Machine Learning Models according to the Four Datasets.

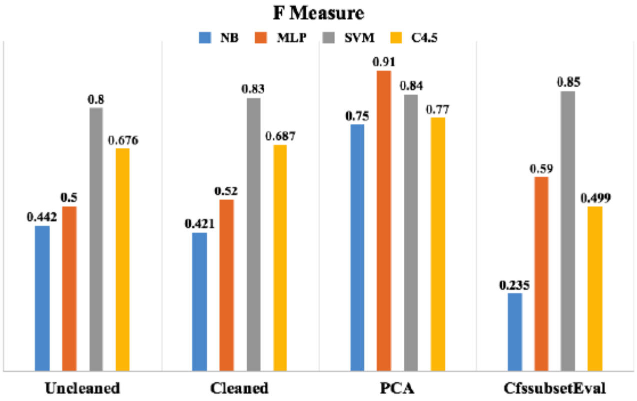


Fig. 7. F Mesure of Machine Learning Models according to the Four Datasets.

The highest rate of accuracy is obtained on MLP predictive model, using dataset with PCA features reduction method. It reached 91% of accuracy.

This presents the MLP model as the most efficient predictive model of semiconductor failures, which can predict future events to avoid failures.

In order to confirm this results, we implemented the MLP model on the python environment, then we obtained 91, 95% of accuracy.

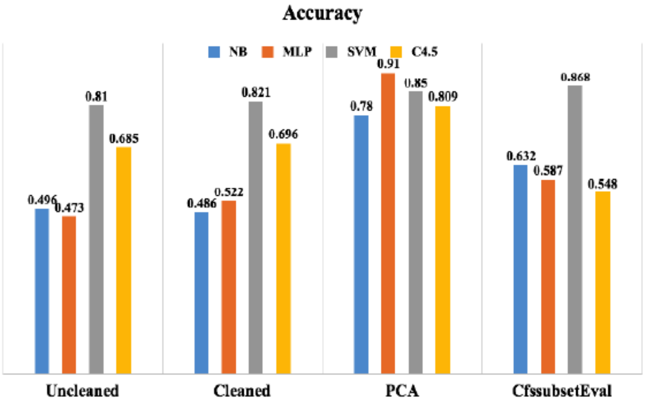


Fig. 8. Accuracy of Machine Learning Models according to the Four Datasets.

V. Conclusion

In order to create a new efficient predictive model of semiconductor failures based on machine learning, we designed and implemented four models based on the most used machine learning algorithms in this field, using SECOM dataset. Due to imbalance records of the success and failure examples in addition to the large amount of data we have proposed in the first part of preprocessing an oversampling method and during the cleaning phase, we have removed the attributes containing a unique value and the average 55% of missing values, and among these remaining features, we selected the most relevant features using CfssubsetEval methods and PCA method.

We performed a series of experiments from which we created four predictive models based on four machine learning algorithms (NB, SVM, MLP, C4.5). We implemented every model on four datasets (uncleaned dataset, cleaned dataset, and cleaned dataset with attributes selected by the CfssubsetEval method, cleaned and reduced with the PCA method.) Five metrics are used for efficiency evaluation (TP Rate, FP rate, F-Measure, Precision, and accuracy). Then, we developed the MLP predictive model on the python environment. The results shows that our predictive model is more efficient and performant, reached 91, 95% of accuracy. We report that data clearance and at-tribute reduction are critical steps in the data-mining process. Therefore we cannot ignore these phases, they require a considerable attention.

It is important to point that for dealing with maintenance events, PdM emerges as an efficient tool. With the Industry 4.0, PdM became gradually very promising. The employment of ML algorithms, for designing PdM applications leads to performant results with cost reduction of a PdM strategy in a factory.

In future works, we aim to use large and complex dataset with various labels, from different equipment on the factory, on real time, In order to identify other relevant features that impact the production line. Also, implement our model on real factory, and shows results on real time.

References

[1] J A. Shohin, X. Xun, L. Yuqian, et al. IoT-enabled smart appliances under industry 4.0: A case study. Advanced Engineering Informatics, 2020, vol. 43, p. 101043.

[2] T. Borgi, A. Hidri,B. Neef and M.S. Naceur. Data analytics for predictive maintenance of industrial robots. Interna-tional conference on advanced systems and electric technologies (IC_ASET) 2017 (pp. 412–417). IEEE.

[3] E. Rauch, C. Linder and P. Dallasega. Anthropocentric perspective of production before and within Indus-try 4.0. Computers & Industrial Engineering, 2020, vol. 139, p. 105644.

[4] T.P. Carvalho, F.A Soares, A. Fabrízzio, R. Vita et al. A systematic literature review of machine learning methods applied to predictive maintenance. Computers & Industrial Engineering, 2019, vol. 137, p. 106024.

[5] S. Biswal & G. R. Sabareesh, Design and development of a wind turbine test rig for condition monitoring studies. 2015 International conference on industrial instrumentation and control, ICIC 2015 (pp. 891–896). IEEE. (2015).

[6] R. S. Peres, R. A. Dionisio, P. Leitao and J. Barata. Idarts - Towards in-telligent data analysis and real-time supervision for industry 4.0. Computers in Industry, 101, 138–146. (2018).

[7] E. Sezer, D. Romero, F. Guedea, M. MacChi and C. Emmanouilidis. An industry 4.0-enabled low cost predictive maintenance approach for SMEs: a use case applied to a cnc turning centre. IEEE International conference on engineering, technology and in- novation (ICE/ITMC) (pp. 1–8). IEEE. (2018).

[8] J. Wan, S. Tang, D. Li, S. Wang, C. Liu, H. Abbas and A. V. Vasilakos. A manufacturing big data solution for active preventive maintenance. IEEE Transactions on Industrial Informatics, 13, 2039–2047. (2017).

[9] G. A. Susto, S. Member, A. Beghi and C. D. Luca. A predictive maintenance system for epitaxy processes based on filtering and prediction techniques. IEEE Transactions on Semiconductor Manufacturing, 25, 638–649. (2012).

[10] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone and A. Beghi. Machine learning for predictive maintenance: A multiple classifier approach. IEEE Transactions on Industrial Informatics, 11, 812–820. (2015).

[11] H. M. Hashemian and W. C. Bean. State-of-the-art predictive maintenance techniques*. IEEE Transactions on Instrumentation and Measurement, 60, 3480–3492. (2011).

[12] G. Biau, and E. Scornet. A random forest guided tour. TEST: An Official Journal of the Spanish Society of Statis-tics and Operations Research, 25, 197–227. (2016).

[13] M. Canizo, E. Onieva, A. Conde, S. Charramendieta and S. Trujillo. Real-time predictive maintenance for wind turbines using Big Data frameworks. IEEE International conference on prognostics and health management (ICPHM) (pp. 8). IEEE. (2017).

[14] A. Kusiak and A. Verma. Prediction of status patterns of wind turbines: A data- mining approach. Journal of Solar Energy Engineering, 133, 1–10. (2011).

[15] C. J. Su and S.F. Huang. Real-time big data analytics for hard disk drive predictive maintenance. Computers and Electrical Engineering, 71, 93–101. (2018).

[16] J. Mathew,M. Luo and C.K. Pang. Regression kernel for prognostics with support vector machines. IEEE Interna-tional conference on emerging technologies and factory automation (ETFA) (pp. 1–5). IEEE. (2017).

[17] T. Wuest, D. Weimer, C. Irgens and K.-D. Thoben. Machine learning in manu- facturing: Advantages, challenges, and applications. Production & Manufacturing Research, 4, 23–45, (2016).

[18] M. Salem, S. Taheri and J. Yuan. An experimental evaluation of fault diagnosis from imbalanced and incomplete data for smart semiconductor manufacturing. Big Data and Cognitive Computing, 2018, vol. 2, no 4, p. 30.

[19] J. Van Hulse and T. Khoshgoftaar."Knowledge discovery from imbalanced and noisy data Data & Knowledge Engineering", 2009.

[20] E.A. Bayrak, P. Kirci and T. Ensari. Performance Analysis of Machine Learning Algorithms and Feature Selection Methods on Hepatitis Disease. International Journal of Multidisciplinary Studies and Innovative Technolo-gies, 2019, vol. 3, no 2, p. 135-138.

[21] I.T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2016, vol. 374, no 2065, p. 20150202.

[22] I. Portugal, P. Alencar, and D. Cowan. The use of machine learning algorithms in recommender systems: A systematic review. Expert Systems with Applications, 2018, vol. 97, p. 205-227.

[23] D. Van Ravenzwaaij and J.P.A. Ioannidis. True and false positive rates for different criteria of evaluating statistical evidence from clinical trials. BMC medical research methodology, 2019, vol. 19, no 1, p. 218.

[24] S. Corbett-Davies and S. Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023, 2018.

Industrial Energy Load Profile Forecasting under Enhanced Time of Use Tariff (ETOU) using Artificial Neural Network

Mohamad Fani Sulaima[1], Siti Aishah Abu Hanipah[2], Nur Rafiqah Abdul Razif[3]
Intan Azmira Wan Abdul Razak[4], Aida Fazliana Abdul Kadir[5], Zul Hasrizal Bohari[6]
Faculty of Electrical Engineering
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia

***Abstract*—The demand response program involves consumers to mitigate peak demand and reducing global CO2 emission. In sustaining this effort, energy provider such as Tenaga Nasional Berhad (TNB) in Peninsular Malaysia has introduced Enhance Time of Use (ETOU) tariff. However, since 2015, small numbers join the ETOU program due to less confidence in managing their energy consumption profile. Thus, this study provides an optimum forecasting load profile model for TOU and ETOU tariffs using Artificial Neural Network (ANN). An industry's average energy profile has been used as a case study, while the forecasting technique has been conducted to find the optimum energy load profile congruently. The load shifting technique has been adopted under ETOU tariff price while integrating to the ANN procedure. A significant comparison in terms of cost reduction between TOU and ETOU electricity tariffs has been made. In contrast, ANN performance results in searching for the best-shifted load profile have been analyzed accordingly. From the proposed method, the total electricity cost saving has been founded to be saved for about 7.9% monthly. It is hoped that this work will benefit the energy authority and consumers in future action, respectively.**

Keywords—Time of use; artificial neural network; energy forecasting; load profile

I. Introduction

Demand Side Management (DSM) consists of Demand Response (DR) program to promote a better independent load management strategy for the consumers in dealing with the pricing and the time allocation. Under the DR framework, there is a price-based program related to time-based price. Under this structure, the consumers will select a program correlated to their energy load profile management. At the same time, they can shift the most available load to the lower price rate at a particular time zone [1]. One of the commands price-based programs is the Time of Use (TOU) tariff.

The TOU tariff has been proposed with various designs and depending on the national policy implemented through energy providers' tariffs to the consumers. The authors in [2] have highlighted the TOU time zones by the region. For example, India's command design was four zones, and China was five-zones and Brazil with three zones. However, the reference [3] presents a different opinion where the arrangement of the time zones design in TOU should be flexible enough to comply with two parties' needs: generation and consumer sides, respectively.

The TOU method design considering both consumers and generation was presented in [4], [5]. The founding is reliable, where the price elasticity consideration has been given attention. The marginal cost of the generation comes to the minimum profit rate is secured. Simultaneously, the promotion of the consumers' participants to the TOU program was a critical factor that was always being discussed. The TOU tariff design is also close to the residential consumers where the behavior and the appliances' arrangement to involve in load management is the command issue. Optimal load arrangement has been practiced while reflecting the energy provider or retailers' TOU tariff as explained in [6]. The consumers' response is essential for the TOU tariff's sustainable design, where the constitution of the demand response program toward human factor contribution is crucially needed.

In Peninsular Malaysia, is facing dramatic changes in the Malaysia Energy Supply Industry (MESI), electricity companies need to examine associated business models and a host of potential strategies to solve equations for over electricity demand from consumers especially in next MESI 2.0 [7]. TNB introduced many schemes, including ETOU tariffs, to benefit the commercial and industrial consumers, but less of them could implement load shifting and join the program correctly.

Thus, in this paper, the analysis and suggestion for the ETOU tariff study to deal with industrial load profile is explored. Meanwhile the implementation of the artificial algorithm has been given attention to produce the optimal profile of the energy to meet the lower cost of the electricity under ETOU program. The related works have been presented in Section II. Meanwhile, the detailed explanation of the specific industrial load profile formulation under the ETOU tariff price has been written in Section III. Meanwhile, Section IV introduced ANN's implementation in processing the load profile forecasting under the ETOU tariff price. The simulation results comparing the actual load profile and the optimization load profile are demonstrated in Section V. The last Section VI concludes the overall study contribution and recommendation for an available future research opportunity.

II. Related Works

Unsuitable tariffs for different load profiles will increase energy costs and lead to the DR program's wrong perception [8]. In [9], [10], the ETOU tariff price has been adopted to the commercial and industrial consumers' reference energy profile. The finding has been analyzed where the flexibility of the load shifting weightage has been constructed accordingly. There were limitations found that the minimum load shifting weightage was higher for the consumers to gain the cost reduction.

The previous study's suggestion contributes to implementing the optimization algorithms to find the optimal load profile that reflects the ETOU tariff price focusing on the Malaysian electricity market condition. The application of the Particle Swarm Optimization (PSO) algorithm has been presented in reference [11] to find the appropriate load categories under six groups of the sectors in peninsular Malaysia. The powerful load management strategies to reduce total electricity under ETOU tariff have been decreased by using Ant Colony Optimization (ACO) [12] and [13]. The authors in [2] present an Evolutionary Algorithm (EA) as the optimal search of load profile in tackling the ETOU price. The future recommendation has been made in the field of knowledge where the optimization algorithm's impact would be further studied to load profile forecasting under the ETOU tariff price.

The application of a conventional vector machine in determining the load forecasting was critically discussed in [14]. Combining the optimization algorithm produced more impact in reducing the root means square value while updating the convergence time, as proven by [15]. As different studies reported of the application of ANN-Self Organizing Mapping for the load forecasting, the significant analysis made has contributed to the excellent summary of the group clustering for the electricity profiles as the example in [16]–[18]. However, those references less reflect the tariff structure in considering the TOU's impact on benefiting consumers of the electricity cost reduction concurrently.

Regarding the load profile forecasting, which reflects ETOU electricity tariff in peninsular Malaysia, there is no particular study to date for the best of the knowledge. Thus, in this study, the mathematical equation stage for ETOU price has been explored and explained. In contrast, the effective formulation for optimal load profile forecasting and the ANN's implementation has been proposed congruently. The significant case study of an industry load profile from the medium voltage category was chosen with the operation time was 24 hours. In contrast, the optimal load profile results to reduce the electricity cost have been explained accordingly.

III. Formulation of Industrial Energy Load Forecasting under ETOU Tariff Price

The flow of the ETOU mathematical stage for demand pricing is presented as follows: Based on the average load profile data, the range for peak time, off-peak time, and maximum demand are identified. Maximum demand cost, C_{TMD}:

$$C_{TMD} = P_{MD} \times R_{TMD} \quad ; (MYR) \tag{1}$$

The average load profile versus time (energy consumption) for every peak are plotted. The reference formula was obtained from the graph by performing regression, and data analysis processes have been shown in the results. Remodel the reference formula by substituting the total peak cost and rated peak value to find a new peak and off-peak ratio. Thus, the total off-peak cost is written as Y_{TOP}:

$$Y_{TOP} = P_{TOP} \times R_{TOP} \quad ; (MYR) \tag{2}$$

Total peak cost, Y_{TP}:

$$Y_{TP} = P_{TP} \times R_{TP} \quad ; (MYR) \tag{3}$$

The formulas for the total cost for every peak are designed by using new ratio values. Verify the total formula cost for peak and off-peak by plotting new graph time versus power using every peak formula's total cost. Perform data analysis and regression to get the formula from the graph. Compared the ratio value with the calculated ratio value from (4). The total cost for TOU tariff, T_T is:

$$T_T = C_{TOP} \times C_{TP} \times C_{TMD} \quad ; (MYR) \tag{4}$$

Where;

C_{TMD} = total electricity cost for maximum demand

C_{TOP} = total electricity cost for off-peak time

C_{TP} = total electricity cost for peak time

Repeat all Eq. (1) until (4) for ETOU tariff, but the arrangement of the mid-peak cost follows Eq. (3). Thus, the total cost for ETOU tariff, T_E is:

$$T_E = C_{EOP} \times C_{EMP} \times C_{EP} \times C_{TMD} \quad ; (MYR) \tag{5}$$

Where;

C_{TMD} = total electricity cost for maximum demand

C_{EOP} = total electricity cost for off-peak time

C_{EMP} = total electricity cost for medium peak time

C_{EP} = total electricity cost for peak time

The T_T and T_E formula is tested with a conventional method by substituting all the ratio, Q and tariff rates, R values, and formula value. To observed TOU and ETOU tariff, graph tariff versus cost TOU and ETOU are performed and analyzed.

Hence, the optimal formulation for daily energy consumption cost (EC) is Eq. (6).

$$EC = \sum_{t=1}^{48} R(t)L(t) \tag{6}$$

Where; R(t) = ETOU rate of each hour

L(t) = Load of each 30 minutes ($0 < t \leq 48$)

Since the ETOU tariff price is fixed every day, consumers may shift their loads to reduce energy costs. Then the load shifting reflect the energy cost, EC_R could be written as

$$EC_R = \sum_{t=1}^{48} R(t)L_R(t) \tag{7}$$

Where; EC_R = Energy cost of that day under load shifting;

$L_R(t)$ = Load shifting of each 30 minutes

$(0 < t \leq 48)$;

The industrial consumers' objective function to enjoy minimum energy cost by load shifting is written in Eq. (8) accordingly.

$$min(EC_R) = \min[\sum_{t=1}^{48} R(t) L_R(t)] \quad (8)$$

The total energy consumption in a day for baseload or actual load and after load shifting must be equal has been set as the constraint. The consumers may only change some of their loads from the high price periods to the low-price periods. But the total consumption maybe not changed due to similar behaviors in their daily life or production. However, in the condition of the simulation forecasting process, if the load shifting condition less than 5% from the actual load, the results can also be considered for comparison. Thus, the constrain condition could be written as:

$$\sum_{t=1}^{48} L_R(t) \approx \sum_{t=1}^{48} L(t) \quad (9)$$

IV. Implementation of Artificial Neural Network

Table I shows stages that are involved in load forecasting and optimization by using ANN. Meanwhile, the explanation of the implementation has been written in sub-section A until E accordingly.

TABLE I. ANN PROCESS TO DETERMINE THE OPTIMAL LOAD PROFILE UNDER ETOU TARIFF PRICE

Stages	Process
Step 1	Data Organization
Step 2	Data Training
Step 3	Data testing
Step 4	Data Forecasting

A. Data Organization

The previous energy profile data is used as input data. The data include daily energy profile value in minutes in January 2016. Twenty-eight data sets are divided into two main groups, which are x_(LoadProfile) and y_EnergyCost. An x_(LoadProfile) is industry energy profile for 14 data set for working days. Meanwhile, y_EnergyCost is for total energy cost per day with ETOU tariff for 14 data set for working days.

B. Data Training

In the training process, the data structure must be normalized. The normalization is copied to the map structure during the trained ANN. The data input and output will normalize and scale the variable values between negative and positive ones. The optimum number of neurons must be considered during normalization. The data are normalized by using the following formula:

$$x_n = \left[x_j - \left(\frac{x_{max}+x_{min}}{2}\right)\right] \div \left(\frac{x_{max}-x_{min}}{2}\right) \quad (10)$$

Where; X_n = Normalized data;

X_j = Actual data;

X_{max} = Maximum value in actual data;

X_{min} = Minimum value in actual data;

C. Data Testing

After training, the maps are tested with testing data of energy profile and energy cost. This method is to associate the energy profile and tariff simultaneously. The data testing process will obtain the pattern for the most suitable load shifting profile.

D. Data Forecasting

After testing, all the load data will be forecasted. The curve of energy profile forecasting and actual profile is obtained by the simulation process accordingly.

V. Results and Analysis

Fig. 1 shows the energy profile, which is having different power consumes during every time zone. The industrial electricity installation is registered under E2 tariff while comparison has been made to E1 flat and E1 ETOU tariff price. The energy pattern for January 25, 2016 (Thursday-Working day) showed that the energy was fully consumed during peak time zones. The Red dot on the figure indicates that the max power consumes 3,330kW, which will be considered maximum demand. The maximum demand position is at 3:00 PM, which is in the peak time zone for TOU and ETOU tariffs. Fig. 2 shows the energy consumed profile for an industrial within 14 days (working day). The profiles demonstrate power consumption for working days is just the same, which is fully consumed during normal working hours, and only 50% operate at night. For the normal 24 hours operation in the manufacturing batch process, the repeating profile would be expected. The load profile index would be calculated as well to see the level of the significant correlation of the maximum demand and the energy consumption.

The best prediction of the minimum energy cost pattern was produced in Fig. 3. The blue line refers to the actual energy cost with ETOU tariff, while the red line shows the new energy cost after optimization with a regression of 0.7003. The graph prediction graph showed various changing patterns to reach the minimum energy cost. Since the ETOU tariff has been divided by Six-Segmentation of the zone that reflect three prices unit, the investigation could be analyzed based on that points. The simulation power profile for the ETOU tariff on the mid-peak and peak zones have followed the baseline profile. However, the consumption of the electricity has reduced significantly. The power consumption on off-peak zone from the 10:00PM to the 8:00AM has increased tremendously to show the performance of the algorithm to find the optimal load to be shifted from peak to off-peak zone. Thus, all those condition during the adaptation of the ANN has contributed to the cost profile such presented in Fig. 3 accordingly.

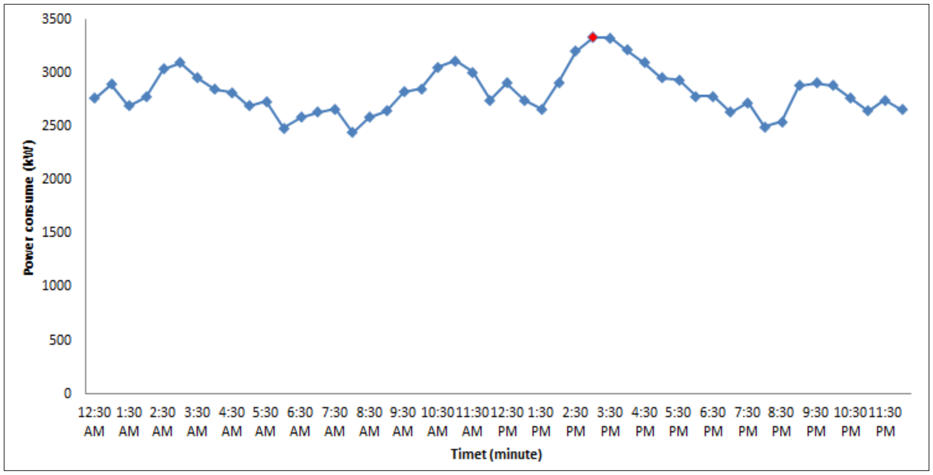


Fig. 1. Energy Profile in every 30 minutes for 24 hours from 12:30 am until 12:00am (average).

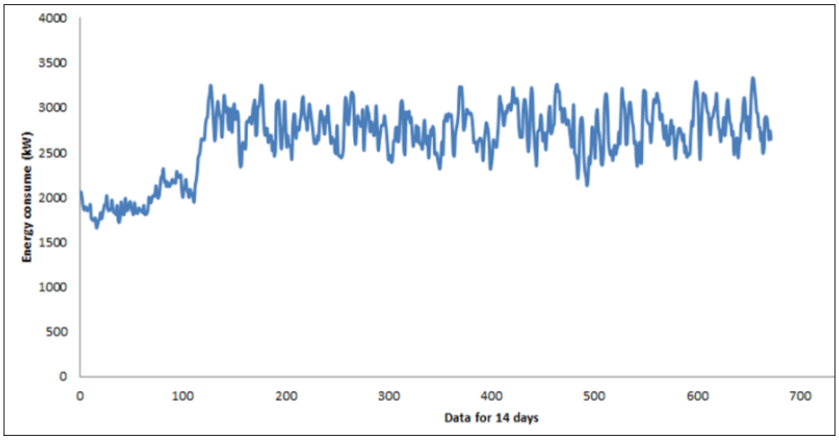


Fig. 2. The Energy Pattern of Industrial Sector for 14 days.

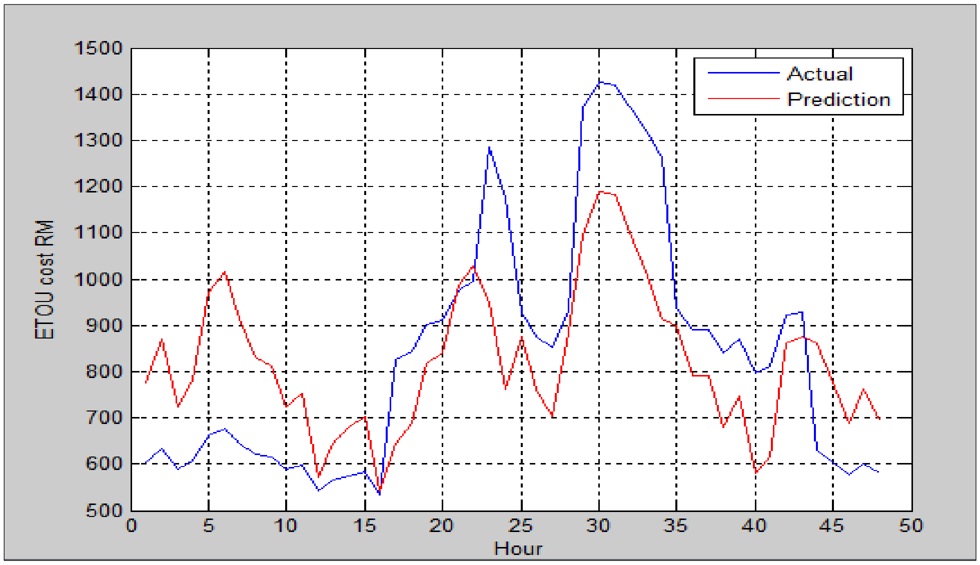


Fig. 3. The Energy Cost Pattern for Actual Data and Prediction Data.

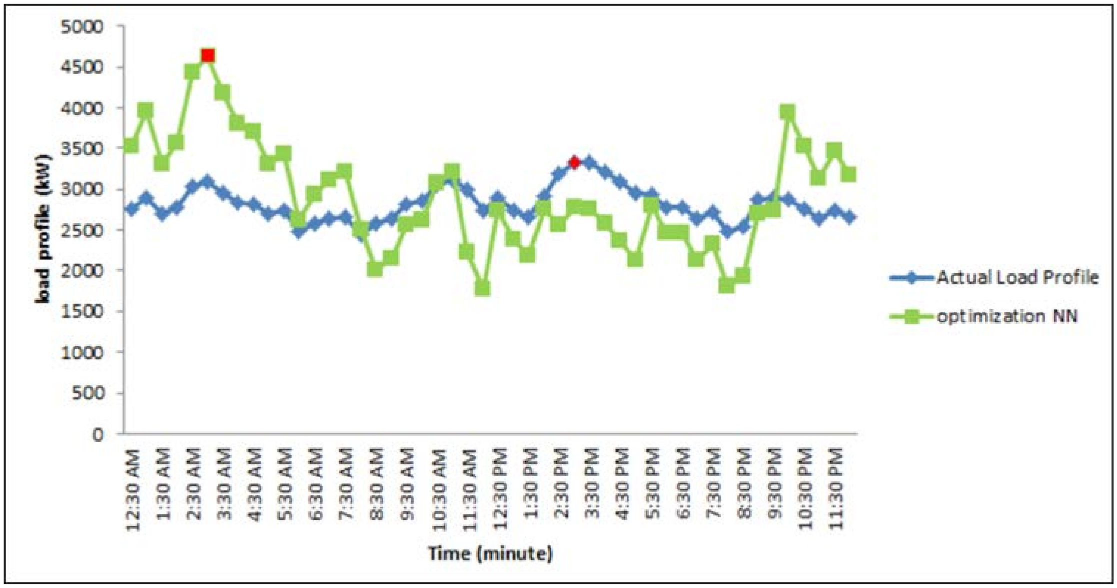


Fig. 4. The Energy Profile Pattern for Actual Data and Optimization Data.

The optimal load response of typical industrial load to ETOU power price for demand-side management are shown in Fig. 4. The customer shifts some of the loads from the high price period to the low-price period in order to achieve minimum energy cost in the day. The load is reduced about 10% at the price peak (about 3:00 PM), and 15% reduction at mid-peak (about 11:30 AM). Meanwhile, the load is increased by about 20% at the price low peak (about 3:00 AM) and 15% improved at low peak (about 10:00PM). Overall load optimization in average ratio was 40:60 of peak and off-peak zone, respectively.

The maximum demand position also shifted from peak region to mid peak region while the peak demand was allocated in off-peak concurrently. It is indicated that the peak electricity consumption is reduced, and the off-peak electricity consumption increases significantly, which affects the power system's regular operation. Table II shows the reduction of the analysis for the energy and maximum demand cost based on two types of tariffs offered to the consumer. Since the ETOU offers off-peak price for the weekend, the advantage has gained to the industrial profile when running the same operation every day. Total forecasting monthly electricity bill when using the optimization method was reduced up to 7.9% which approximately MYR 103,453.00 or yearly saving for about MYR 1,241,436.00.

Besides, Table III presents the comparison of the cost reduction between the case of optimum TOU tariff and case of the optimum ETOU tariff price accordingly. Both of the price rates has been used in the simulation. Meanwhile, the optimum load profile was defined as explained before. Through the comparison, both TOU and ETOU are produced excellence cost reduction for approximately 5%~8%, but it would be recommended that the ETOU tariff scheme be able to benefit the consumers with the condition that the optimum load must be applied. The study's cost-saving recorded shows the significant contribution of the demand response program to the consumers. In contrast, the peak demand would be projected to be shifted to off-peak hour. Thus, the peak demand movement condition contributes to effort the efficiency management of the generation side where the critical type of the generation would not be run in the long time every day.

TABLE II. THE COST COMPARISON BETWEEN ACTUAL ENERGY PROFILE AND OPTIMIZATION ENERGY PROFILE

Records	**Actual Case E1 flat**	**Optimization Case E1 ETOU**	**Reduction**
Daily Energy cost	40,201.00	39,369.52	2.07%
Maximum demand (kW)	3,330 (peak)	3,210 (mid peak)	3.60%
Maximum demand cost (MYR)	98,568.00	95,016.00	3.60%
Total Monthly Electricity Cost (MYR)	1,304,598.00	1,201,145.44	7.90%

TABLE III. THE COMPARISON OF ENERGY COST BY USING OPTIMIZATION ENERGY PROFILE

Tariff	Without Optimization Monthly Energy Cost, (MYR)	With Optimization Monthly Energy Cost, (MYR)	Saving (%)
TOU	1,262,019.00	1,194,424.80	5.36
ETOU	1,318,238.00	1,201,145.44	8.80

VI. CONCLUSION

In this study, the ANN has been applied to forecast the optimum load profile under TOU and ETOU tariff prices. The comparison of the both tariffs has been done where the benefit of the TOU and ETOU price could be received by consumers were quantified based on the monthly and yearly results. Meanwhile, the significant is shifted load must converge in the zone of the mid-peak for the lower charge of the maximum demand. As the investigation statement presents that the ETOU can be optimum adopted by the consumers where the minimum load shifting condition is applied. Besides, the electricity bill's cost can be reduced; the peak demand has been shifted to the off-peak zone that brings benefits to the generation side. The future recommendation of the research continuity can be explored on the comparison of the forecasting algorithms by using the other techniques while using a multi-objective environment. Meanwhile other factor of the effect toward load factor index and the energy economic side should be considered as well.

ACKNOWLEDGMENT

The authors would like to thank Universiti Teknikal Malaysia Melaka (UTeM) for all the support given. The publication of the paper is sponsored under the publication support scheme under Centre for Research and Innovation UTeM (CRIM).

REFERENCES

[1] F. Meyabadi and M. H. Deihimi, “A review of demand-side management: Reconsidering theoretical framework,” Renew. Sustain. Energy Rev., vol. 80, no. January 2016, pp. 367–379, 2017.

[2] M. F. Sulaima, N. Y. Dahlan, Z. M. Yasin, N. A. M. Asari, and Z. H. Bohari, “Optimum enhance time of use (ETOU) for demand side electricity pricing in regulated market: An implementation using evolutionary algorithm,” Indones. J. Electr. Eng. Comput. Sci., vol. 8, no. 1, 2017.

[3] V. Venizelou, N. Philippou, M. Hadjipanayi, G. Makrides, V. Efthymiou, and G. E. Georghiou, “Development of a novel time-of-use tariff algorithm for residential prosumer price-based demand side management,” Energy, vol. 142, pp. 633–646, 2018.

[4] M. Pauzi Abdullah, M. N. Nazatul Shiema, M. Y. Hassan, and F. Hussin, “Optimizing Time of Use Electricity Pricing in Regulalted Market,” J. Teknol., vol. 7, pp. 31–39, 2013.

[5] N. S. M. Nazar, M. P. Abdullah, M. Y. Hassan, and F. Hussin, “Time-based electricity pricing for Demand Response implementation in monopolized electricity market,” in SCOReD 2012 - 2012 IEEE Student Conference on Research and Development, 2012, pp. 178–181.

[6] S. Yilmaz, S. Weber, and M. K. Patel, “Who is sensitive to DSM ? Understanding the determinants of the shape of electricity load curves and demand shifting : Socio-demographic characteristics, appliance use and attitudes,” Energy Policy, vol. 133, pp. 1–13, 2019.

[7] Energy Commission (Malaysia), “Reimaging Malaysia Energy Supply Industry (MESI 2.0),” 2019.

[8] S. Mohajeryami, I. N. Moghaddam, M. Doostan, B. Vatani, and P. Schwarz, “A novel economic model for price-based demand response,” Electr. Power Syst. Res., vol. 135, pp. 1–9, 2016.

[9] N. A. M. Azman, M. Pauzi Abdullah, M. Y. Hassan, and D. M. Said, “Enhanced Time of Use Electricity Pricing for Industrial Customers in Malaysia,” Indones. J. Electr. Eng. Comput. Sci., vol. 6, no. 1, pp. 155–160, 2017.

[10] N. A. M. Azman, M. P. Abdullah, M. Y. Hassan, D. M. Said, and F. Hussin, “Enhanced time of use electricity pricing for commercial customers in Malaysia,” Pertanika J. Sci. Technol., vol. 25, pp. 285–294, 2017.

[11] M. F. Sulaima, N. Y. Dahlan, Z. M. Yasin, M. M. Rosli, Z. Omar, and M. Y. Hassan, “A review of electricity pricing in peninsular Malaysia: Empirical investigation about the appropriateness of Enhanced Time of Use (ETOU) electricity tariff,” Renew. Sustain. Energy Rev., vol. 110, pp. 348–367, 2019.

[12] M. Sulaima, N. Dahlan, and Z. Yasin, “Effective Electricity Cost Management in a Manufacturing Operation by Using Optimal ETOU Tariff Formulation,” Int. J. Electr. Electron. Syst. Res., vol. 15, pp. 82–93, 2019.

[13] M. F. Sulaima, N. Y. Dahlan, M. H. Isa, M. N. Othman, Z. M. Yasin, and H. A. Kasdirin, “ETOU electricity tariff for manufacturing load shifting strategy using ACO algorithm,” Bull. Electr. Enginnering Informatics, vol. 8, no. 1, pp. 21–29, 2019.

[14] D. Niu, Y. Wang, and D. Dash, “Power load forecasting using support vector machine and ant colony optimization,” Expert Syst. Appl., vol. 37, pp. 2531–2539, 2010.

[15] A. Ghasemi, H. Shayeghi, M. Moradzadeh, and M. Nooshyar, “A novel hybrid algorithm for electricity price and load forecasting in smart grids with demand-side management,” Appl. Energy, vol. 177, pp. 40–59, 2016.

[16] N. . Z. N. F. . G. N.N Atira; I. Azmiara, Z.H. Bohari, “Medium Term Load Forecasting Using Statistical Feature Self Organizing Maps (SOM),” J. Telecommun. Electron. Comput. Eng., vol. 11, no. 2, pp. 25–29, 2019.

[17] X. Lin, Z. Tian, Y. Lu, H. Zhang, and J. Niu, “Short-term forecast model of cooling load using load component disaggregation,” Appl. Therm. Eng., vol. 157, pp. 1–13, 2019.

[18] Z. H. Bohari, H. S. Azemy, M. N. M. Nasir, M. F. Baharom, and M. F. Sulaima, “Reliable Short Term Load Forecasting Using Self Organizing Map (SOM) in Deregulated Electricity Market,” J. Theor. Appl. Inf. Technol., vol. 79, no. 3, pp. 389–394, 2015.

A Fast Military Object Recognition using Extreme Learning Approach on CNN

Hari Surrisyad[1]
Master Program in Computer Science
Universitas Gadjah Mada, Yogyakarta, Indonesia

Wahyono[2]*
Department of Computer Science and Electronics
Gadjah Mada University, Indonesia

***Abstract*—Convolutional Neural Network (CNN) is an algorithm that can classify image data with very high accuracy but requires a long training time so that the required resources are quite large. One of the causes of the long training time is the existence of a backpropagation-based classification layer, which uses a slow gradient-based algorithm to perform learning, and all parameters on the network are determined iteratively. This paper proposes a combination of CNN and Extreme Learning Machine (ELM) to overcome these problems. Combination process is carried out using a convolution extraction layer on CNN, which then combines it with the classification layer using the ELM method. ELM method is Single Hidden Layer Feedforward Neural Networks (SLFNs) which was created to overcome traditional CNN's weaknesses, especially in terms of training speed of feedforward neural networks. The combination of CNN and ELM is expected to produce a model that has a faster training time, so that its resource usage can be smaller, but maintaining the accuracy as much as standard CNN. In the experiment, the military object classification problem was implemented, and it achieves smaller resources as much as 400 MB on GPU comparing to standard CNN.**

Keywords—Training-speed; resource; backpropagationm; CNN; ELM

I. Introduction

In recent years, the field of computer vision has been developed to support advanced systems in various fields such as intelligent robots, automatic control systems, and human-computer interaction. On the other hand, one of the applications in the military field is automatic target detection, which is the main technology for automatic military operations and surveillance missions [1]. Military objects are legitimate targets for attack in war [2].

Convolutional Neural Network (CNN) is a popular algorithm that excels in vector data classification, which belongs to deep learning algorithms. CNN is a special type of neural network that handles phenomena such as localization of receptive fields in large data volumes, copying weights forward, as well as image sampling using different kernels in each convolution layer [3]. Convolution is a process in which an image is manipulated by using an external mask to produce a new image [4]. CNN uses a feedforward neural network with backpropagation-based learning at its classification layer or what is often called the fully connected layer. The feedforward neural network has the disadvantage of using a slow gradient-based learning algorithm for learning [5]. All parameters on the feedforward neural network must be determined manually iteratively, the parameters in question are the input weight and hidden bias. These parameters are also interconnected between layers, so they are often stuck on the local optima and require a long learning time and lots of resources.

Extreme Learning Machine (ELM) is a feedforward neural network with a single hidden layer or commonly known as Single Hidden Layer Feedforward Neural Networks (SLFNs). The ELM learning method can overcome weaknesses of CNN, especially in terms of rapid training of the feedforward neural network [5]. Therefore, a combination of convolutional neural networks and extreme machine learning is proposed, by replacing the backpropagation method used at the CNN classification layer with the ELM method which can overcome the weakness of backpropagation. This combination is expected to increase learning speed become faster so that the utilization of resources during training is getting smaller, but with accuracy the same as for regular CNN.

This research is expected to be used in situations where a system, especially in the military field, requires small resources and prioritizes speed. An example of a real implementation that can be done is in a surveillance drone, which can recognize military objects. Therefore, drones and adjust the distance to the recognized objects. In this case, the military objects can be recognized from far distance, such as military aircraft, military helicopters, and others, as well as at close range such as grenades, pistols, rifles, and so on.

II. Related Work

Many researches related to the introduction of military objects with CNN and ELM have been carried out. One of them is deep transfer learning for military object Recognition under small training set condition [6]. This research focuses on the classification and recognition of objects with a limited amount of data with CNN, with transfer learning to provide knowledge and combining various layers to perform better feature extraction. It obtained an average value of 95% accuracy. Another method is recognizing military vehicles in social media images using deep learning [7]. The research was evaluated using dataset which was collected from various social media, namely Flickr, YouTube, and the Web. In the experiment, it achieved an accuracy of 95.18%. However, both method still requires slow processing time and large resource in training step.

ELM is a feedforward neural network with a single hidden layer or commonly called a single hidden layer feed-forward

*Corresponding Author

neural network, which has advantages in learning speed. One of the ELM utilization is for recognizing facial expression which was done by Mahmud and Al Mamun [8]. In this research, facial expression image recognition was classified into six classes, with ELM and Backpropagation Neural Network as a comparison, ELM obtained an accuracy of 90% and backpropagation 86%, while the ELM speed was 0.0936 second and backpropagation 1 second with a total of 42 image data. Another utilization of ELM is proposed by Wiyono which implemented ELM as classifier for face recognition combining with PCA for feature extraction [9]. By using JAFFE Dataset, the method obtained an accuracy of 93.1% with a training speed of 0.062 seconds.

Research combining CNN with other algorithms is also nothing new, one of them is Convolutional SVM Networks for Object Detection in UAV Imagery proposed by Bazi and Melgani [10]. In this research, the network used is based on several alternatives convolutional and a reduction layer, which was then combined with the SVM as classification layer. This is done in order to obtain an optimal model for classification and prediction, with very limited training data. The resulting accuracy in this research is 97% for the Car dataset and 96% for the solar panel dataset.

III. Methodology

A. Research Goal

In this research, we aim to overcome the weaknesses of backpropagation used in convolutional neural networks. It is expected that the proposed method could increase the speed of training step so that the resources used are also getting smaller. The flow of our proposed method is shown in Fig. 1.

B. Data Acquisition

In this research, we collected military object image data which consists of 16 different classes, with 15 military object classes and 1 non-military object class. The data was collected from Google images, using Google-images-download library, this library is made with the Python programming language. It is then divided into training and testing data, along with the object image class to be used. Fig. 2 shows several selected samples of military object images collected in our dataset.

C. Data Preprocessing

Data preprocessing is a series of processes carried out on data so that the data is ready to be used as input in the training process. There are many types of image data preprocessing that can be done. In this research, the preprocessing that will be carried out is as follows:

1) Data cleaning: After the data is obtained during the acquisition process, the data will be cleaned first. The cleaning process is carried out by deleting data that does not match the criteria as follows: (1) data in the form of weapons that are not being held or used, (2) vehicle data, taken from the side or tilt angle that represents the shape of the vehicle in general, (3) data according to the object class that has been defined, and (4) the image data only containing one object, except the army object class.

2) Data augmentation: The next preprocessing is data augmentation. The data augmentation is required because the number of data obtained in the dataset is very limited, such as only 350 data per class. For obtaining a good classification accuracy, the data should be large enough. However, it would be very difficult to collect such large data manually, so that we employ the data augmentation for increasing the number of data. Data augmentation example is shown in Fig. 3.

3) Resizing: The next step is resizing the image data. This resizing process is carried out to equalize the image size of each data, because the data obtained from Google images have various sizes and dimensions. In this research, image data will be resized to 224 × 224 pixels similar to our input layer size on the CNN architecture. The illustration of resizing process is depicted in Fig. 4.

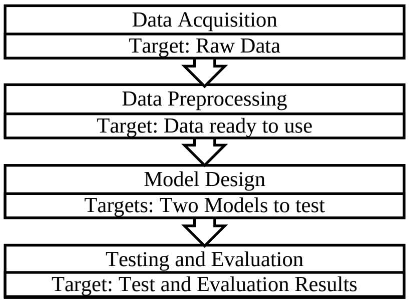

Data Acquisition
Target: Raw Data
Data Preprocessing
Target: Data ready to use
Model Design
Targets: Two Models to test
Testing and Evaluation
Target: Test and Evaluation Results

Fig. 1. Research Flow.

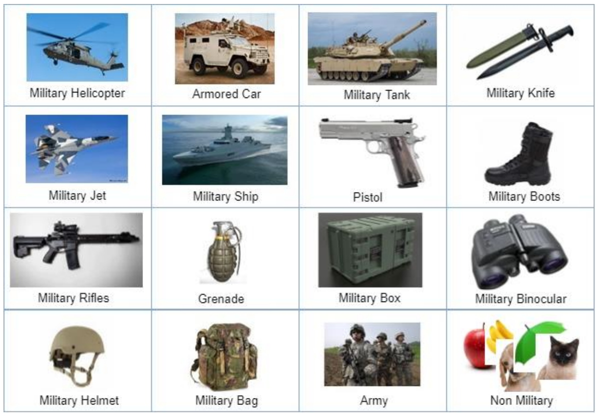
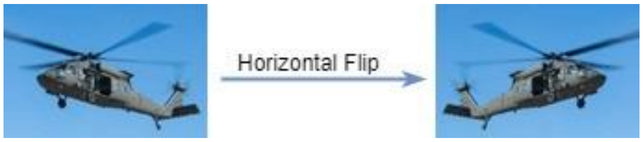


Fig. 2. Sample of Military Object Images used in our Proposed Dataset.



(a) (b)

Fig. 3. Data Augmentation Ilustration using Horizontal Flip (a) Raw Data (b) Augmentation Results.

(a) (b)

Fig. 4. Resize the Image to 224 × 224.

D. Model Design

In this research, after the data is ready, a model design process will be carried out to perform the learning process of the training data from each class. The designed model will greatly affect the classification results.

1) Normal CNN: CNN is a convolutional operation that combines multiple layers of processing, uses several elements operating in parallel and is inspired by the biological nervous system [11]. In CNN, each neuron is represented in two dimensions, so this method is suitable for processing with input in the form of images [12]. The CNN structure consists of input, feature extraction process, classification process and output. The extraction process on CNN consists of several hidden layers, namely the convolution layer, the activation function (ReLU), and pooling, as shown in Fig. 5.

In designing the CNN model, there are many types of architectures that can be made. Each architecture with certain data must go through a tuning process to get a model that is considered optimal. Fig. 6 is the initial architecture that will be used for further tuning in this research.

The tuning process is carried out by making gradual changes to the initial architecture that has been determined. There are many parameters that can be set in the CNN model such as number of convolution and concatenation operations, order of each operation, kernel in convolution and concatenation operations, number of hidden layers in FCL, number of nodes in each hidden layer and many more. Tuning process will be stopped if the optimal model has been found.

2) Combination of CNN and ELM: ELM is a feedforward neural network with a single hidden layer or commonly called Single Hidden Layer Feedforward Neural Networks (SLFNs) which only requires two parameters, namely the number of hidden nodes and the choice of activation function. The ELM learning method is designed to overcome the weaknesses of the feedforward neural network, especially in terms of learning speed. Based on two reasons why feedforward ANN has a slow learning speed:

- Using slow gradient based learning algorithms for conducting training.
- All parameters on the network are determined iteratively using this learning method.

In ELM, parameters such as input weights and hidden bias are chosen randomly, so that ELM has the ability to learn quickly and is able to produce good generalization performance.

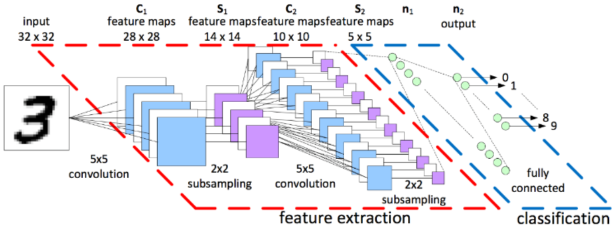


Fig. 5. CNN Architecture Baseline [13].

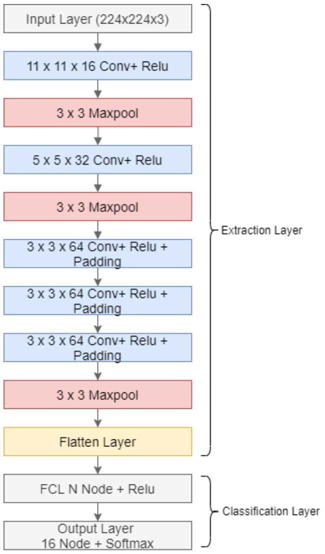


Fig. 6. Initials CNN Architecture.

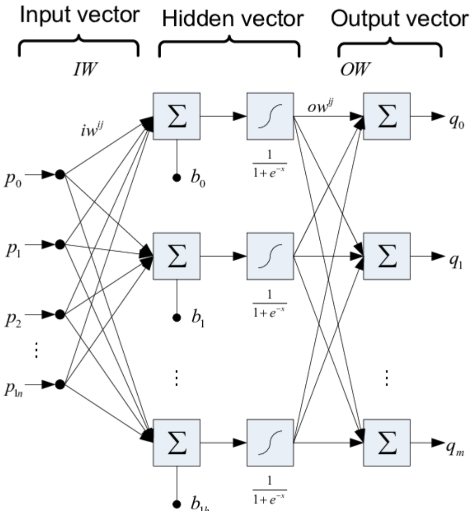


Fig. 7. ELM Network [14].

The ELM method has a different mathematical model from the feedforward neural network, as shown in Fig. 7. The ELM mathematical model is simpler and more effective. For N different number of input pairs and output targets (x_i, t_i), with $x_i = [x_{i1}, x_{i2}, . . . , x_{in}]^T \in \boldsymbol{R}^n$ and $t_i = [t_{i1}, t_{i2}, . . . , t_{in}]T \in \boldsymbol{R}^m$, Standard SLFN with the number of hidden nodes and the activation function $g\ (x)$ can be modeled mathematically as follows:

$$\sum_{i=1}^{\tilde{N}} \beta_i g_i(x_j) = \sum_{i=1}^{\tilde{N}} \beta_i g(w_i \,.\, x_j + b_i) = o_j \,, j \ = \ 1, 2, \dots, N \quad (1)$$

where:

$w_i = [w_{i1}, w_{i2}, ..., w_{in}]^T$ is a weight vector that connects *hidden node* i and *input nodes.*

$\beta_i = [\beta_{i1}, \beta_{i2}, ..., \beta_{im}]^T$ is the connecting weight vector *hidden node* i and *output nodes.*

bi is threshold from hidden node i

w_i . x_j is *inner product* from w_i and x_j

Standard SLFNs with $\tilde{N}$ hidden nodes and activation function $g(x)$ assumed to be able to estimate N of this sample with an error rate of 0 which means $\sum_{j=1}^{N}\left\|o_j - t_j\right\| = 0$, so there is β_i, w_i, and b_i that:

$$\sum_{i=1}^{\tilde{N}} \beta_i g\left(w_i . x_j + b_i\right) = t_j , j = 1,2, \dots , N \tag{2}$$

The above equation can be simply written as:

$$H\beta = T \tag{3}$$

where:

$$H = \begin{bmatrix} g(w_1 . x_1 + b_1) & \cdots & g(w_{\tilde{N}} . x_1 + b_{\tilde{N}}) \\ \vdots & \ddots & \vdots \\ g(w_1 . x_n + b_1) & \cdots & g(w_{\tilde{N}} . x_N + b_{\tilde{N}}) \end{bmatrix},$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_{\tilde{N}}^T \end{bmatrix} and\ T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}$$

H in the above equation is the hidden layer output matrix of the neural network. $g(w_i . x_j + b_i)$ shows the output of hidden neurons related to input xj. β is the output weight matrix and T is the target matrix. In ELM, the input weight and hidden bias are determined randomly, so that the output weight associated with the hidden layer can be determined from the equation:

$$\beta = H^+ T \tag{4}$$

In the equation above H^+ is the *Moore-Penrose Generalized invers* matrix of the H matrix. H^+ is obtained by the equation:

$$H^+ = (H^T .H)^{-1}.H^T \tag{5}$$

H is the hidden layer output matrix and H^T is the transpose of H. Following are the steps in the Extreme Learning Machine (ELM) algorithm:

Input : input pattern x_j and target output pattern t_j, j = 1, 2,..N

Output: input weight w_i, output weight β_i and *bias* b_i, i = 1,2...$\tilde{N}$

Steps :

1: Determine the activation function (g (x)) and the number of hidden nodes ($\tilde{N}$).

2: Determine the random value of the input weight w_i and *bias* b_i, i = 1, 2, ..., $\tilde{N}$.

3: Calculate the output matrix value H on the hidden layer.

4: Calculate the output weight value β using $\beta = H^+ T$.

5: Calculate the output value with $H\beta = T$.

In this research, the combination layer feature extraction model of CNN and ELM will use the same layer as the feature extraction layer in the normal CNN model that has been tuned. The difference is that this combination model classification layer will replace the FCL which uses backpropagation as the basis for learning with ELM.

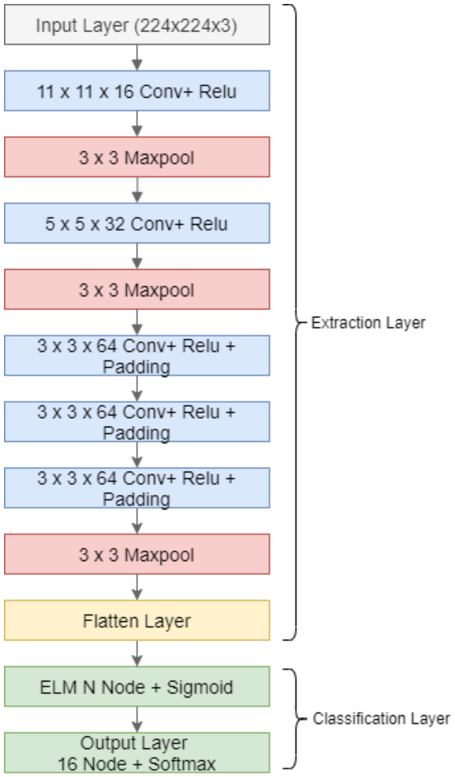


Fig. 8. Initial Combined Architecture of CNN and ELM.

The ELM classification layer will be tuned again. However, the only parameters that will be tuned are the number of nodes in the hidden layer and their activation function. On the other hand, the number of hidden layers will not be set because basically ELM is a single hidden layer feedforward neural network (SLFNs), as shown in Fig. 8.

E. Testing and Evaluation Design

After the model design process is complete, two different models will be obtained, namely normal CNN and Combination of CNN and ELM. Furthermore, several testing and evaluation processes will be carried out. Fig. 9 is the test and evaluation design scheme that will be carried out in this research.

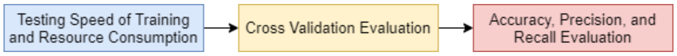

Testing Speed of Training and Resource Consumption
Cross Validation Evaluation
Accuracy, Precision, and Recall Evaluation

Fig. 9. Testing and Evaluation Design.

1) Testing speed of training and resource usage: In the testing process, the two methods will be implemented and then calculated how long it will take for the training time, this measurement will be done in seconds. Testing the use of resources required by both methods, the resource referred to here is the use of memory during the training process.

Testing of training speed and resource usage is conducted on several factors, such as, the amount of data, the variation of the extraction layer, the number of hidden layers (FCL classification layer), and the number of hidden layer nodes (classification layer). The detail comparison schema for training speed and resources is shown in Table I.

TABLE I. TRAINING SPEED TESTING SCHEME AND RESOURCE USAGE

Model	Factor	Testing Speed	Testing Resources
Normal CNN	The amount of data	√	√
	Extraction layer variations	√	√
	Number of hidden layers (classification layer)	√	√
	Number of hidden layer nodes (classification layer)	√	√
Proposed Combination of CNN and ELM	The amount of data	√	√
	Extraction layer variations	√	√
	Number of hidden layers (classification layer)	×	×
	Number of hidden layer nodes (classification layer)	√	√

2) Cross validation evaluation: Furthermore, the cross-validation evaluation process will be carried out on the two models that have been made. Cross-validation was carried out to evaluate the accuracy of the two models that have been made against the training data.

All training data will be divided into n subsets evenly with the same size, then the training and testing process is carried out n times repeatedly. In iteration 1, subset 1 becomes validation data and the other becomes training data. In iteration 2, subsets 2 becomes validation data and others become training data, and so on until it has been finished, as shown in Fig. 10.

3) Accuracy, precision, and recall evaluation: The final evaluation that will be carried out is the evaluation process on the test data. This process is applied by classifying all test data which also has 15 classes. These data are data that are not included in the training process. Then the process of calculating accuracy, precision and recall will be carried out using following equations*:*

$$Accuracy = \frac{\sum_{i=1}^{l}\frac{TP_i+TN_i}{TP_i+TN_i+FP_i+FN_i}}{l} * 100\% \quad (6)$$

$$Precision_{class} = \frac{TP_{class}}{TP_{class}+FP_{class}} * 100\% \quad (7)$$

$$Recall_{class} = \frac{TP_{class}}{TP_{class}+ FN_{class}} * 100\% \quad (8)$$

where

- TP_i is *True Positive*, that is, the number of positive data classified correctly by the system for class i.
- TN_i is *True Negative*, that is, the number of negative data classified correctly by the system for class i.
- FN_i is *False Negative*, that is, the amount of negative data but incorrectly classified by the system for class i.
- FP_i is *False Positive*, that is, the number of positive data but incorrectly classified by the system for class i.
- l is number of classes.

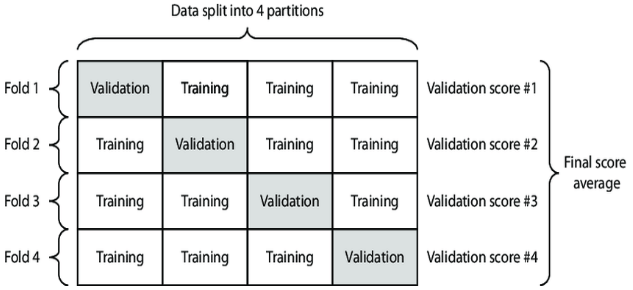


Fig. 10. Illustration of 4-Fold Cross Validation.

IV. EXPERIMENTS RESULTS

A. Data Acquisition Results

The data acquisition process that has been carried out using the *google_images_download* library with various keywords in each object class, has succeeded in collecting 16 classes of raw data with different amounts of data in each class. The results of data acquisition can be seen in the Fig. 11.

B. Results of Data Preprocessing

The raw data that has been collected will then go through several preprocessing stages, this is done to compile the raw data into data that is ready for use, and several processes are carried out as follows:

1) Cleaning data: The first step is the cleaning process. This process is carried out to clean data that is incompatible with existing classes. The result of this process is data that contains and is in accordance with the existing class, in each class 350 images are selected, so that the total data in the data set is 5,600 images.

2) Resizing: The next process is resizing data. All the data that have been selected have a very diverse size. To simplify the modeling process, all data will be equalized in pixel size to 224 × 224 pixels. An example of the results of the resizing process can be seen in Fig. 12.

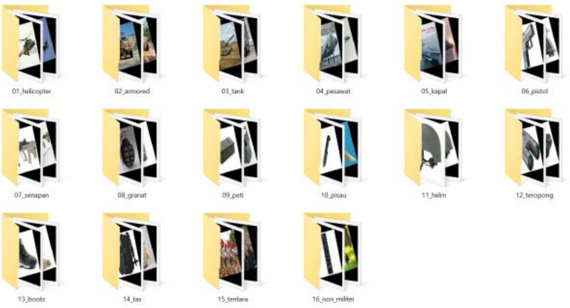


Fig. 11. Data Acquisition Results.

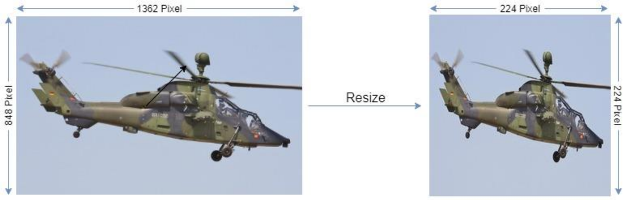


Fig. 12. Resizing Data Results.

3) Augmentation data: The next process is data augmentation. This process is conducted to increase the amount of data, so that the model has enough data. Therefore, it can be used for the training process and produces a good model. Some of the data augmentation used are as follows:

- Flip Horizontal

The first augmentation is a horizontal flip. This process is performed to flip the image horizontally. In the result of this process, the data is duplicated, so that each class will have a total of 700 data, and the total data in the dataset is 11,200. An example of the results of the flipping process can be seen in the Fig. 13.

- Rotating

The second augmentation is rotating. This process is applied to rotate the image, in this research the image is rotated. The result of this process the data is increased threefold, each class the number becomes 1,050 data, so that the total data in the dataset is 16,800a. This data will be used in the modeling process. An example of the result of the rotation process can be seen in the Fig. 14.

- Shifting

The third augmentation is shifting. This process is performed to shift the position of the pixels in the image. In this research the pixels are shifted 30 pixels to the right. The result of this process the data is increased fourfold, each class the number becomes 1,400 data so that the total data is 22,400 data. This data will be used in the testing process. An example of the results of the shifting process can be seen in the Fig. 15.

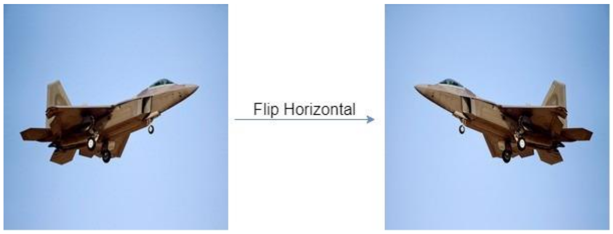


Fig. 13. Results of the Flipping Process.

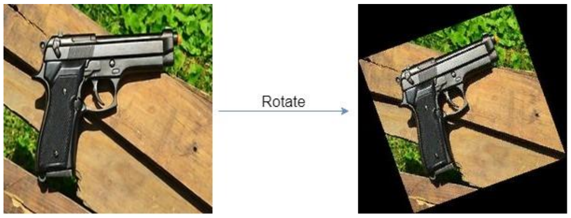


Fig. 14. Results of the Rotation Process.

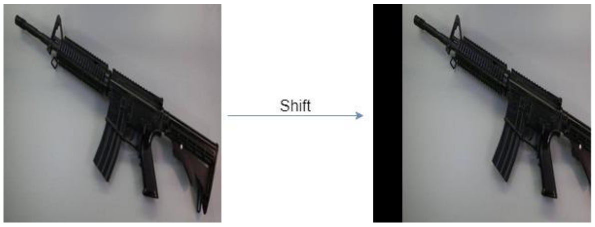


Fig. 15. Result of the Shift Process.

C. Modeling Results

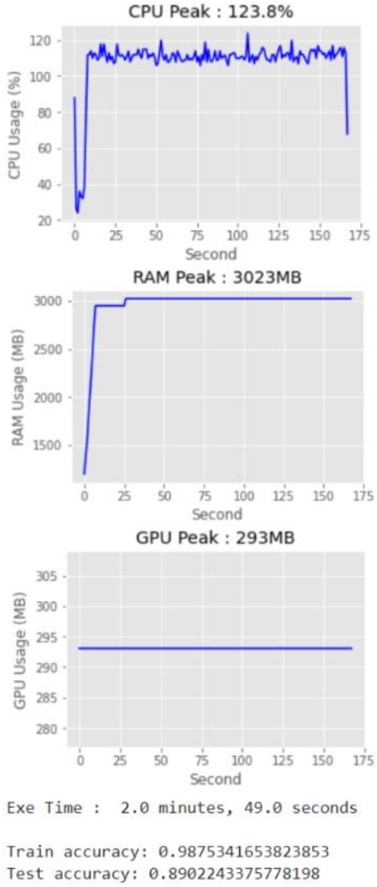
Data that is ready and has a sufficient amount will be used in the modeling process. The data used in the modeling process is 1,050 per class and a total of 16,800 data as a whole. The data will be divided into training and testing data with a ratio of 80:20. After conducting experiment, the modeling results are as follows.

1) Normal CNN model: The first modeling process is Normal CNN, with the initial architecture that has been determined at the beginning of the research. The results of the training are as shown in Fig. 16.

In the training process above, the training time is 2 minutes 49 seconds, with peak resource usage of 123.8% CPU, 3032 MB RAM, and 293 MB GPU. In the training process, it obtains accuracy of 0.987, while the data test was 0.890.

From the initial architecture, the tuning process is conducted. After going through a long, we obtain the optimal architecture as shown in Fig. 17.

Fig. 18 shows the results of training from tuned CNN architectures. In the training process with a tuned architecture, the training time is 4 minutes 30 seconds, with peak resource usage, such as CPU 156.8%, RAM 3300 MB, and GPU 771 MB. It obtains the training accuracy of 0.984 while the test data was 0.924.



```
Exe Time :  2.0 minutes, 49.0 seconds

Train accuracy: 0.9875341653823853
Test accuracy: 0.8902243375778198
```

Fig. 16. Initial Normal CNN Architecture Training Results.

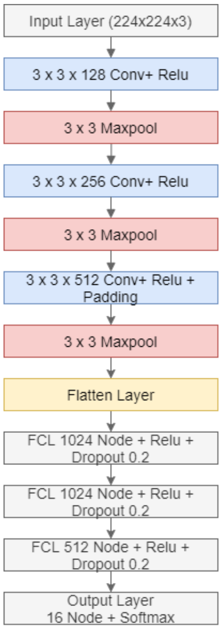
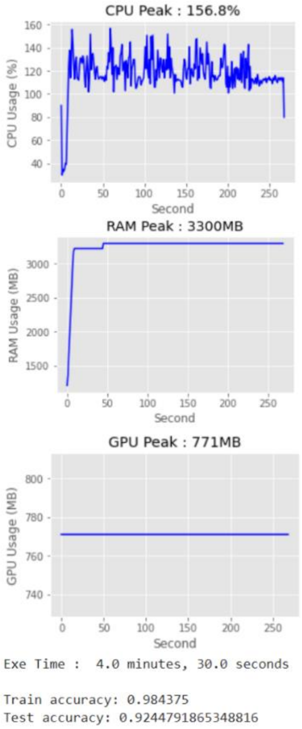


Fig. 17. Optimal CNN Architecture Obtaining by Tuning Process.



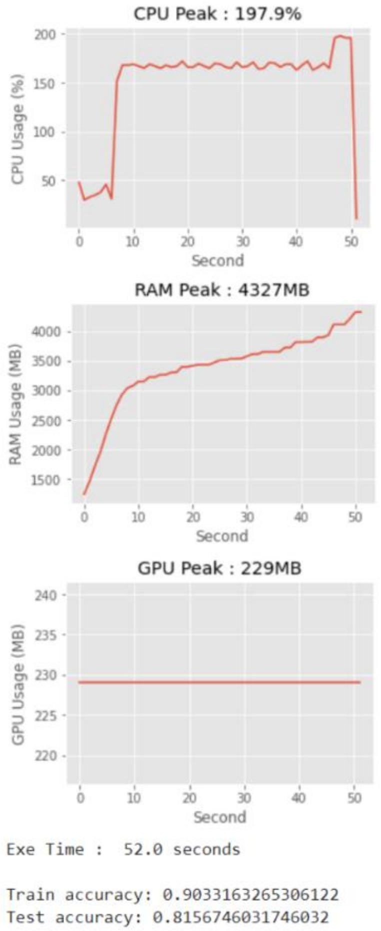
```
Exe Time :  4.0 minutes, 30.0 seconds

Train accuracy: 0.984375
Test accuracy: 0.9244791865348816
```

Fig. 18. The Tuned CNN Normal Architecture Training Results.

2) Combination of CNN and ELM model: The next modeling process is the combination modeling of CNN and ELM. Using the initial architecture, the training results are shown in Fig. 19.

In the training process of the combined CNN and ELM model with the initial architecture, the training speed is 52 seconds, with peak resource usage of 197.9% CPU, 4327 MB RAM, and 229 MB GPU. The accuracy in training was 0.903 while the test data is 0.815.

The combined CNN and ELM model also goes through a tuning process, and the tuning results are shown in Fig. 20. Fig. 21 shows the results of training from Combined Architecture of CNN and ELM.



```
Exe Time :  52.0 seconds

Train accuracy: 0.9033163265306122
Test accuracy: 0.8156746031746032
```

Fig. 19. Results of Initial Architectural Training for a Combination of CNN and ELM.

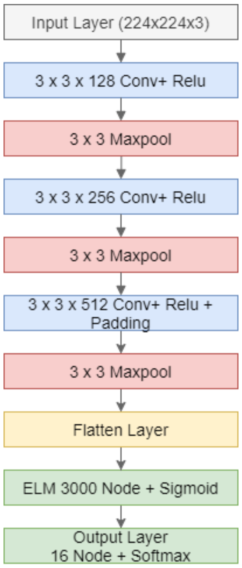
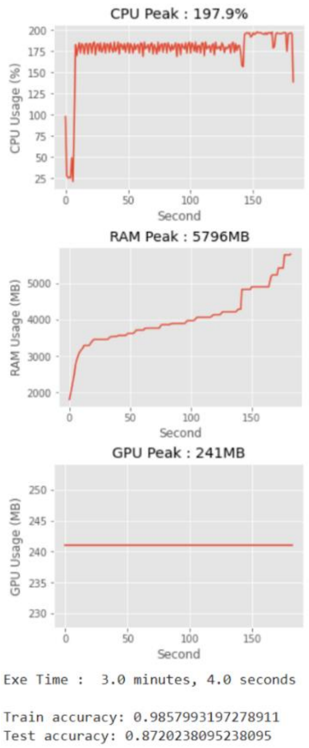


Fig. 20. Combined Architecture of CNN and ELM after Tuning Process.



```
Exe Time :  3.0 minutes, 4.0 seconds

Train accuracy: 0.9857993197278911
Test accuracy: 0.8720238095238095
```

Fig. 21. Results of Tuned Architecture Training from Combination of CNN and ELM.

In the architecture that has been tuned the training process above, the training time is 3 minutes 4 seconds, with peak resource usage of 197.9% CPU, 5796 MB RAM, and 241 MB GPU. In training, we obtain accuracy of 0.985 while the test data was 0.872.

D. Testing and Evaluation Results

The model that has been made in the previous process will be tested with a test scenario that has been made, with several aspects and factors, to find out how well the model is performing.

1) Testing training speed and resource usage: In this test, the model will be tested on how long training time and how large resource use are associated with accuracy, with the following factors:

- The amount of data

This factor is tested to determine how much influence the amount of data has on the training process, by increasing the amount of data from 1,050 per class to 1,400 data per class so that the total data becomes 22,400.

- Variation of the Extraction Layer

In this factor, tests are carried out to determine how much influence the complexity of the extraction layer has on the training process. At this stage an additional layer of convolutional extraction is added to the architecture.

- Number of hidden layers

This factor is tested to determine how much influence the number of hidden layer classification on the training process, on normal CNN plus one hidden layer. In the combination model of CNN and ELM, this stage is not carried out because ELM only has one hidden layer.

- The number of hidden layer nodes

This factor is tested to determine how much influence the number of hidden layer nodes has on the classification process of the training process. In normal CNN the third hidden layer is increased from 512 to 1024 nodes. For the combination of CNN and ELM model hidden nodes increased from 2500 to 300 nodes. After conduting experiment using above factors, the results of this process can be seen in Table II.

2) Cross validation evaluation: The next scenario is evaluation with the cross-validation method. This process is carried out to evaluate the accuracy of the two models that have been made against the training data. This research will use 5-fold cross validation, which means that the training data will be divided into five parts. This evaluation is shown in Table III.

The results above, when plotted with the line chart, are shown in Fig. 22.

TABLE II. RESULTS OF TESTING TRAINING SPEED AND RESOURCE USAGE

Model	Factor	*Training Time*	Resource Usage (Peak)	Accuracy
Normal CNN	Amount of data	6 minutes 3 seconds	CPU 158.9%, RAM 3233MB, GPU 771MB	Train: 0.97 Test: 0.89
	Variation layer extraction	2 minutes 57 seconds	CPU 118.9%, RAM 2662MB, GPU 432MB	Train: 0.96 Test: 0.88
	Number of hidden layers	4 minutes 29 seconds	CPU 153.9%, RAM 3301MB, GPU 771MB	Train: 0.97 Test: 0.91
	Number of hidden layer nodes	6 minutes 2 seconds	CPU 140.9%, RAM 2483MB, GPU 753MB	Train: 0.96 Test: 0.89
Proposed Combination of CNN and ELM	Amount of data	4 minutes 14 seconds	CPU 197.9%, RAM 7074MB, GPU 259MB	Train: 0.97 Test: 0.86
	Variation layer extraction	1 minutes 41 seconds	CPU 197.9%, RAM 5753MB, GPU 259MB	Train: 0.98 Test: 0.85
	Number of hidden layer nodes	3 minutes 49 seconds	CPU 197.9%, RAM 6255MB, GPU 241MB	Train: 0.98 Test: 0.86

TABLE III. RESULTS 5-FOLD CROSS VALIDATION OF NORMAL CNN

Iteration	Accuracy
Iteration 1	0.87
Iteration 2	0.89
Iteration 3	0.90
Iteration 4	0.88
Iteration 5	0.90
Average	0.89

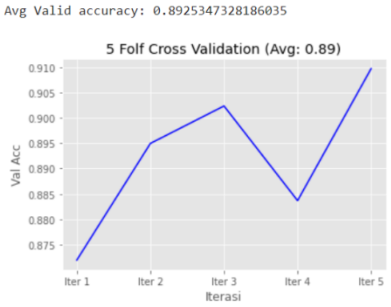


Fig. 22. Plot of Results 5-Fold Cross Validation Normal CNN.

The results of the evaluation of the combined CNN and ELM models can be seen in the following Table IV. The results, when plotted with the line chart, are shown in Fig. 23.

3) Accuracy, precision, and recall evaluation: The last scenario is the evaluation of accuracy, precision, and recall of data testing using confusion matrix, this is done to find out how well the model can generalize knowledge.

In the normal CNN model the results of confusion matrix can be seen in the following Fig. 24.

From confusion matrix, accuracy, precision, and recall can be calculated. The results can be seen in the following Table V.

In Table V, the precision value is obtained with a Micro Average of 0.92 and an Average Macro of 0.92. On the other hand, the recall value with a Micro Average of 0.92 and an Average Macro of 0.92.

Average micro calculates the metric independently for each class and then takes the average, suitable for cases with a balanced amount of data for each class. Whereas Average Macro represents the contribution of all classes as whole to calculate the metric mean, it is suitable for cases with a balanced amount of data.

For the combination of CNN and ELM model, the results of the confusion matrix can be seen in the Fig. 25.

From confusion matrix, accuracy, precision, and recall can be calculated, the results of which can be seen in the following Table VI.

In the Table VI, the precision value obtained with Avg Micro is 0.88 and Avg Macro is 0.88. On the other hand, the recall value with Avg Micro was 0.88 and Avg Macro was 0.88.

TABLE IV. RESULTS OF 5-FOLD CROSS VALIDATION COMBINATION OF CNN AND ELM

Iteration	Accuracy
Iteration 1	0.86
Iteration 2	0.85
Iteration 3	0.87
Iteration 4	0.85
Iteration 5	0.86
Average	0.86

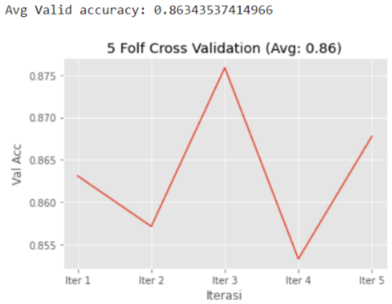


Fig. 23. Plot of Result 5-Fold Cross Validation Combination of CNN and ELM.

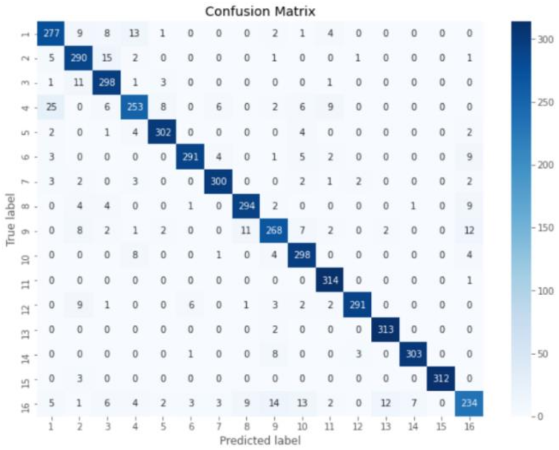
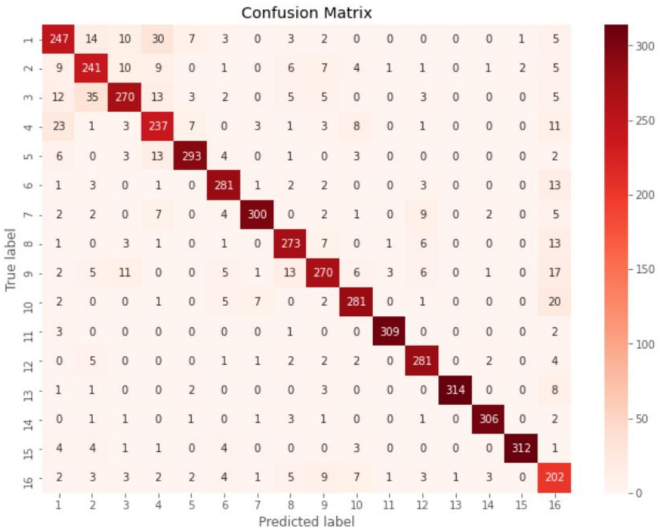


Fig. 24. Confusion Matrix Normal CNN Model.

TABLE V. TABLE NORMAL CNN ACCURACY, PRECISION, AND RECALL RESULTS

Accuracy	0.92	
Class	Precision	Recall
Military Helicopter	0.86	0.88
Armored Car	0.86	0.92
Military Tank	0.87	0.95
Military Jet	0.88	0.80
Military Ship	0.95	0.96
Pistol	0.96	0.92
Military Rifle	0.96	0.95
Grenade	0.93	0.93
Military Box	0.87	0.85
Military Knife	0.88	0.95
Military Helmet	0.93	1.00
Military Binoculars	0.98	0.92
Military Boot	0.96	0.99
Military Bag	0.97	0.96
Army	1.00	0.99
Non-Military	0.85	0.74
Avg Micro	0.92	0.92
Avg Macro	0.92	0.92



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	247	14	10	30	7	3	0	3	2	0	0	0	0	0	1	5
2	9	241	10	9	0	1	0	6	7	4	1	1	0	1	2	5
3	12	35	270	13	3	2	0	5	5	0	0	3	0	0	0	5
4	23	1	3	237	7	0	3	1	3	8	0	1	0	0	0	11
5	6	0	3	13	293	4	0	1	0	3	0	0	0	0	0	2
6	1	3	0	1	0	281	1	2	2	0	0	3	0	0	0	13
7	2	2	0	7	0	4	300	0	2	1	0	9	0	2	0	5
8	1	0	3	1	0	1	0	273	7	0	1	6	0	0	0	13
9	2	5	11	0	0	5	1	13	270	6	3	6	0	1	0	17
10	2	0	0	1	0	5	7	0	2	281	0	1	0	0	0	20
11	3	0	0	0	0	0	0	1	0	0	309	0	0	0	0	2
12	0	5	0	0	0	1	1	2	2	2	0	281	0	2	0	4
13	1	1	0	0	2	0	0	0	3	0	0	0	314	0	0	8
14	0	1	1	0	1	0	1	3	1	0	0	1	0	306	0	2
15	4	4	1	1	0	4	0	0	0	3	0	0	0	0	312	1
16	2	3	3	2	2	4	1	5	9	7	1	3	1	3	0	202

Fig. 25. Confusion Matrix Combination of CNN and ELM Model.

TABLE VI. TABLE COMBINATION OF CNN AND ELM ACCURACY, PRECISION, AND RECALL RESULTS

Accuracy	**0.87**	
Kelas	**Precision**	**Recall**
Military Helicopter	0.78	0.77
Armored Car	0.77	0.81
Military Tank	0.86	0.76
Military Jet	0.75	0.80
Military Ship	0.93	0.90
Pistol	0.89	0.92
Military Rifle	0.95	0.90
Grenade	0.87	0.89
Military Box	0.86	0.79
Military Knife	0.89	0.88
Military Helmet	0.98	0.98
Military Binoculars	0.89	0.94
Military Boot	1.00	0.95
Military Bag	0.97	0.97
Army	0.99	0.95
Non-Military	0.64	0.81
Avg Micro	**0.88**	**0.88**
Avg Macro	**0.88**	**0.88**

E. Analysis and Discussion

Based on the training results in Table II, in the factor of extraction layer variation, one additional convolutional extraction layer and one max pooling layer are added to the architecture. This factor evaluates how much influence the complexity of the extraction layer has on the training process. It is found that the combination model of CNN and ELM achieves processing time 1 minute 43 seconds, which is faster than the normal CNN model. This is because the addition of the extraction layer affects the number of kernels that must be trained iteratively. The effect is that the learning time in the normal CNN model is getting longer, whereas in the combination model CNN and ELM does not carry out a repetitive weight updating process. Therefore, the number of extraction layers does not really affect the combination model of CNN and ELM. For resource usage, the combination CNN and ELM models use 79% more resources on CPU than normal CNN models. The combined CNN and ELM models use 3091 MB more resources on RAM than normal CNN models. The normal CNN model use 176 MB more resources on GPUs compared to the combined CNN and ELM models, gradually. In the training data, the combined CNN and ELM model has a higher accuracy than 0.01 normal CNN model, while the CNN normal model test data is 0.03 superior to the CNN and ELM combination model.

In the factor of the number of hidden layers, it evaluates how much influence the number of hidden layer classifications has on the training process. In this factor, it is only tested on the normal CNN model because the combination model of CNN and ELM only has one hidden layer. It is found that the leaning time in the Normal CNN model is 2 minutes 20 seconds longer than before the addition of the hidden layer, as well as the previous factor, such as the addition of the number of hidden layers has an effect on the amount of weight that must be trained iteratively. The effect is that the tilt velocity in the normal CNN model is getting slower. For resource usage, CPU has 30.1% more resources than without adding hidden layers, normal CNN model RAM uses 1 MB more resources than without adding hidden layers, on normal CNN GPUs the number of resources is the same as before adding hidden layers. In training and testing data, the normal CNN model has smaller accuracy of 0.01 compared to with CNN without the addition of a hidden layer.

In the number of hidden layer nodes factor, it evaluates how much influence the number of hidden layer nodes has on the classification process of the training process. In the third normal CNN, hidden layer is increasing from 512 to 1024 nodes. On the other hand, in the combination model CNN and ELM, hidden nodes are increased from 3000 to 3500 nodes. It is found that the combined model of CNN and ELM require processing time as long as 2 minutes 13 seconds faster than the normal CNN model. This is because the increase in the number of nodes affects the number of weights that must be trained iteratively. Consequently, the leaning time in the normal CNN model is getting longer, while in the combination of CNN and ELM does not perform a repeated weight updating process. Therefore, the number of extraction layers does not really affect the combination model of CNN and ELM. For resource usage, the combination of CNN and ELM models uses 57% more resources on CPU than normal CNN models, combined CNN and ELM models use 3772 MB more resources on RAM than normal CNN models, normal CNN model uses 512 MB more resources on GPUs compared to combined CNN and ELM models. In the training data, the combination of CNN and ELM models have an accuracy of 0.02 which is superior to the normal CNN normal, while the normal CNN model achieve accuracy in test data around 0.03 which is superior to the combination of CNN and ELM models.

The results of the cross-validation evaluation in Tables III and IV show that the average validation accuracy of the normal CNN model is superior, namely 0.89 compared to the average validation accuracy in the combined CNN and ELM model, which is 0.86. It can be seen that both models produce fairly even accuracy. In each part of the cross-validation evaluation process.

For the evaluation of accuracy, precision, and recall, the results are obtained in Tables V and VI. Both from the accuracy, precision and recall of normal CNN models are superior to the combination of CNN and ELM models. This indicates that the normal CNN model has a better generation capability, but with a single layer and without the weight updating process the combination of CNN and ELM has produced very good performance as well. If we look further at the results' confusion matrix on the combination of CNN and ELM model, the prediction error occurs in objects that have many features, helicopters with aircraft and armored cars with tanks. It can be seen that the multilayer FCL on CNN has better ability in the pattern features that are similar or complex compared to a single layer in ELM.

V. Conclusions

From the research process that has been implemented, several conclusions can be drawn as follows:

- The combined CNN and ELM model uses a convolutional extraction layer on CNN, which is then combined with the classification layer using the ELM method. The model learning time is always shorter, approximately 2 minutes, compared to normal CNN. It is because the normal CNN uses full connected layer (FCL) based backpropagation, which still uses slow gradient-based learning algorithms to carry out learning.
- The normal CNN model resource usage is 57% smaller on CPU resources and uses an average of 3568 MB of smaller resources on RAM, but the combined CNN and ELM models uses 400 MB of smaller resources on GPUs.
- Accuracy, precision and recall of normal CNN models are slightly higher by 0.03 to 0.04 compared to combined CNN and ELM models. However, with one layer and without updating process, the combined weight of CNN and ELM was maintaining the accuracy.

References

[1] S. Liu and Z. Liu, “Multi-Channel CNN-based Object Detection for Enhanced Situation Awareness,” pp. 1–9, 2017, [Online]. Available: http://arxiv.org/abs/1712.00075.

[2] E. Prasetiawan, “Implementation of Distinction Principles Related to Civil and Military Object in Indonesia (in Bahasa Indonesia),” Universitas Airlangga, 2019.

[3] M. Sharma, A. Bhave, and R. R. Janghel, “White Blood Cell Classification Using Convolutional Neural Network,” in Soft Computing and Signal Processing, 2019, pp. 135–143.

[4] Y. A. Hambali, “C # Based Process Area Application using Visual Studio (in Bahasa Indonesia),” Ilmu Komput., p. 14, 2011.

[5] G. Bin Huang, Q. Y. Zhu, and C. K. Siew, “Extreme learning machine: A new learning scheme of feedforward neural networks,” IEEE Int. Conf. Neural Networks - Conf. Proc., vol. 2, pp. 985–990, 2004, doi: 10.1109/IJCNN.2004.1380068.

[6] Z. Yang et al., “Deep transfer learning for military object recognition under small training set condition,” Neural Comput. Appl., vol. 31, no. 10, pp. 6469–6478, 2019, doi: 10.1007/s00521-018-3468-3.

[7] T. Hiippala, “Recognizing military vehicles in social media images using deep learning,” 2017 IEEE Int. Conf. Intell. Secur. Informatics Secur. Big Data, ISI 2017, pp. 60–65, 2017, doi: 10.1109/ISI.2017.8004875.

[8] F. Mahmud and M. Al Mamun, “Facial Expression Recognition System Using Extreme Learning Machine,” Int. J. Sci. Eng. Res., vol. 8, no. 3, pp. 266–267, 2017, [Online]. Available: http://www.ijser.org.

[9] A. R. Wiyono, “Introduction to Face Expression Image Using Principal Component Analysis (PCA) and Extreme Learning Machine Algorithm (in Bahasa Indonesia),” Jurnal Ilmiah Matermatika (MATH), vol. 6, no. 2, pp. 2–6, 2018.

[10] Y. Bazi and F. Melgani, “Convolutional SVM Networks for Object Detection in UAV Imagery,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 6, pp. 3107–3118, 2018, doi: 10.1109/TGRS.2018.2790926.

[11] F. Hu, G. S. Xia, J. Hu, and L. Zhang, “Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery,” Remote Sens., vol. 7, no. 11, pp. 14680–14707, 2015, doi: 10.3390/rs71114680.

[12] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, “Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification,” Ieee Tgrs, vol. 55, no. 2, pp. 645–657, 2016, doi: 10.1109/TGRS.2016.2612821.

[13] N. Sharma, V. Jain, and A. Mishra, “An Analysis of Convolutional Neural Networks for Image Classification,” Procedia Comput. Sci., vol. 132, no. Iccids, pp. 377–384, 2018, doi: 10.1016/j.procs.2018.05.198.

[14] L. Deng and D. Yu, “Deep Learning: Methods and Applications,” Found. Trends®in Signal Process., vol. 7, no. 3–4, pp. 197–387, 2014, doi: 10.1561/2000000039.

A New Traffic Distribution Routing Algorithm for Low Level VPNs

Abdelwahed Berguiga[1], Ahlem Harchay[2], Ayman Massaoudi[3], Radhia Khdhir[4]
Department of Computer Science
Jouf University, Sakakah
Saudi Arabia

***Abstract*—Virtual Private Networks (VPN) constitute a particular class of shared networks. In such networks, the resources are shared among several customers. The management of these resources requires a high level of automation to obtain the dynamics necessary for the well-functioning of a VPN. In this paper, we consider the problem of a network operator who owns the physical infrastructure and who wishes to deliver VPN service to his customers. These customers may be Internet Service providers, large corporations and enterprises. We propose a new routing approach referred to as Traffic Split Routing (TSR) which splits the traffic as fairly as possible between the network links. We show that TSR outperforms Shortest Path Routing (SPR) in terms of the number of admitted VPN and in terms of Quality of Service.**

Keywords—Virtual Private Networks (VPN); Quality of Service (QoS); NS-2; Simulations; Shortest Path Routing (SPR); Traffic Split Routing (TSR); Routing algorithm

I. Introduction

With the exponential growth of the Internet and increasingly supports various types of applications, especially those calling on multimedia as well as several users simultaneously, the Internet service provider as well as the network operator are called upon to guarantee commitments of quality of service to their subscribers. The simplicity and low cost of IP networks are some of the reasons why users are deploying new types of applications on these networks. However, some types of real-time applications including video conferencing and VoIP which are very sensitive to variations in delay (jitter) and throughput are not guaranteed with IP.

Reservation of resources for multimedia applications is necessary to ensure end-to-end performance. However, this reservation is not supported with IP. Also, real-time applications also require a guarantee on resources such as storage space, CPU time etc. Thus, the packets must be routed based on the required QoS, which is not possible with the Internet today. However, the Internet is by nature "Best Effort" and lacks any control over the quality of service. Traditional Layer 3 routing methods like Routing Information Protocol (RIP), Open Shortest Path First (OSPF) and Border Gateway Protocol (BGP) would become obsolete if we want to support QoS in the Internet.

A more viable alternative to traditional IP routing and which includes the use of technologies and network infrastructures that guarantee QoS, we can cite IP / MPLS, IP over Metro Ethernet or also IP over ATM. However, even if these technologies support QoS, the topology of the network as well as the routes which will carry the traffic must be correctly chosen otherwise the cost of QoS would be prohibitive. Perhaps the best-known example is the problem of the taxi driver who has to find the quickest and easiest way to get from one place to another. Therefore, instead of letting each driver individually make the decision to choose their route, we instead need to inform them in advance which route they should take. Therefore, the optimal choice of routes must be calculated in advance, i.e. the solution to adopt to guarantee QoS is the "proactive" approach and not the "reactive" approach.

In order to find the optimal routes, we have to define the optimality criteria, i.e. the objectives and the constraints. When a network is given, which is the case with an operator who owns the physical infrastructure, the goal is to increase the number of "satisfied" customers and therefore income. When we begin to build a network, as in the case of a service provider that does not have the infrastructure, the goal is to minimize the use of the leased resources and therefore the cost of the network. In both cases, QoS is a constraint. Moreover, Virtual Private Networks (VPN) constitute a particular class of shared networks. In such networks, the resources are shared among several customers. The management of these resources requires a high level of automation to obtain the dynamics necessary for the well functioning of a VPN.

It is in this context that the present research work “Distributed traffic routing for low-level VPNs” falls within this framework, where we will design in an optimized way VPNs based on logical topologies or multipoint virtual circuits such as such as VPLS (Virtual Private LAN Service), E-LAN (Ethernet LAN Services) etc. We propose to study the case of an operator who has the physical infrastructure and who wants to offer this kind of VPN service to its customers. The operator then seeks to maximize the number of customers while providing the QoS required by each of them.

The rest of the paper is organized as follows: Section II provides an overview on the graph generations using Waxman and Brite algorithms. Section III describes the proposed solution. Section IV reports the performance evaluation environment and the simulation methodology. We report and explain simulation results, useful to assess the validity of our proposed traffic distribution algorithm. Finally, the last section draws conclusions.

II. Related Works

The design of virtual private networks brings together a whole set of optimization problems that differ by the constraints imposed, and sometimes by the data considered. Studying these issues is very important for network operators as well as service providers. The methods of solving these problems often call graph theory, performance analysis and optimization, descriptive as diverse as they are complex.

For a supplier, it is about setting up a network that guarantees the delivery of all its customers' requests with quality of service requirements, while minimizing network operating costs. Operators seek customer satisfaction by trying to make the most of all the resources available in their networks.

A. Random Generation of Graphs

The study of large networks is becoming increasingly important, especially thanks to the evolution of telecommunications networks and the Internet. These networks can be of a different nature. To model them, we often use the formal structure of graphs. A graph is modeled by a set of vertices connected by edges. We can enrich the structure of the graph by assigning a cost to each of the edges. It is often difficult to represent a network accurately. We often prefer to represent the local properties of a network, then we generate the graph while respecting these properties as much as possible. Among other things, the generation of graphs allows us to do simulations.

During this research work, we used a tool that generates random graphs known as “Brite” [1]. Brite allows a random generation of several graphs following the “Waxman” model [2]. The latter offers us the possibility of obtaining large networks whose characteristics resemble those of an Internet network. Once a network is generated, the "Waxman" model assigns two parameters to each link: cost and time.

The principle of the "Waxman" method is as follows [2]:

1) Enter the number of nodes to generate.

2) Calculate the probability P (u, v) of adding a link between each pair of nodes u and v.

$$P(\{u,v\}) = \beta\, e^{\frac{-d(u,v)}{L\alpha}}$$

Where:

$d(u,v)$: the distance between node u and v.

L: the maximum distance between two nodes.

α and β: Two parameters that vary in the interval (0.1];

The increase in β results an increase in the density of links in the graph.

The decrease in α results in an increase in the density of the short links between the nodes.

For each $P(u,v)$, draw a random number T between (0,1]. If T <P, then add a link between u and v.

B. Routing Algorithms

New VPN technologies have greatly expanded the range of possibilities for users. On the one hand, they allow very great flexibility for users, and on the other hand, they lead to increasing complexity for operators and service providers. Several considerations must be examined in order to ensure satisfactory quality of service (QoS) following customer requests. Among these considerations, we can cite:

1) Optimal allocation of communication resources according to user needs and available resources in the network.

2) The establishment of reliability control mechanisms.

On the other hand, the quality of service offered to a connection is directly related to the choice of the path between a source and a destination. The route calculation must take into account the various constraints imposed by a connection (speed, variation in delay, loss rate, etc.). In this outcome, it is necessary to set up a routing algorithm whose role is to find the best path between a source and its recipient while respecting the various constraints imposed. We speak of routing with constraints. Because these constraints vary from customer to customer, and from one type of network to another, it is almost impossible to find a routing algorithm that meets all needs. Indeed, it was proved in [3], that the problem of finding a path with multiple constraints is NP-Complete.

Several heuristic proposals were then presented to solve this problem. These proposals can be classified into five categories:

1) The first approach is to minimize a single QoS parameter. The algorithms of Dijkstra and Bellman-Ford are examples of this approach. They find the shortest route between a source and its destination.

2) The second approach is presented by [4]. An algorithm based on the minimization of a QoS parameter subject to a second constraint is proposed. It uses the cost and the delay calculated by a “distance vector” protocol maintained at each node.

3) The third approach is to build a path under two constraints simultaneously (usually time and cost). Chen et al. [5] and Jaff et al. [6] have proposed algorithms to solve the problem with two constraints. The major problem with this proposal is that it is more complex than the other heuristics and it does not guarantee scalability.

4) The fourth approach is based on minimizing the different parameters in a specific order. The Widest-Shortest and Shortest-Widest [7] algorithms are examples of this approach.

5) The fifth approach to the routing problem with QoS is to construct paths using a combined metric that is calculated based on two (or more) constraints. Verma et al. [8] combined cost and bandwidth into a single metric.

The routing approaches with QoS presented previously vary from the simplest, such as Dijkstra and Bellman-Ford, which are based on a single constraint, to the more complex exploiting two or more constraints. However, these approaches have a common weakness in that they do not guarantee the formation of a balanced system when distributing the load. They use an order of priority in the choice of constraints which leads to the construction of unbalanced paths.

In this research work, we consider the problem of a network operator who owns the physical infrastructure and who wishes to deliver VPN service to his customers. These customers may be Internet Service providers, large corporations and enterprises. We propose a new routing approach referred to as Traffic Split Routing (TSR) which splits the traffic as fairly as possible between the network links. We show that TSR outperforms Shortest Path Routing (SPR) in terms of the number of admitted VPN and in terms of Quality of Service.

III. Proposed Algorithm

In what follows, we will present a simple algorithm, Traffic Split Routing (TSR) [9], having as main objective the load sharing in a network. Indeed, with the use of TSR we will try to distribute the traffic in the network as homogeneously as possible. Our approach is to be able to use the network for a balanced sharing of traffic [10], [11]. Our main goal is to avoid overloading some links while others remain unused. This is often achieved by creating disjointed trees and/or paths and small sizes.

We present in what follows the heuristic of traffic distribution used:

Traffic distribution heuristic

Given ls the number of times a link s appears in a VPN tree.

1: Initialize ls ← 0 for all links.

2: Initialize n ← 0 and wait for a new VPN connection request (or a new site to add to an existing VPN).

3: Generate (or complete) a "generic" tree (path) linking all the new VPN sites without using the links whose ls> n.

4: If the tree is not completed, increment n ←n + 1, and go back to step 3.

Otherwise (the tree is complete), increment ls ←ls + 1 for all the links of the new generated tree and go back to step 2.

First, we define a variable called Link-Usage Count [LUC] (LUC refers to the variable "ls" in the previous algorithm) which gives the number of times a link appears in a VPN tree. This variable will be used as a metric for the generation of trees. In fact, every time a link is used in a tree, its LUC [12] is incremented by one. When generating new trees, our algorithm will try to avoid links with the highest LUC.

Obviously, when a VPN connection ends or one of the sites disconnects, the "ls" value of each link belonging to that connection is decremented by one. The generic tree generated in step 3 can be obtained with any algorithm or protocol. For example, we can use the minimum weight tree (MST) [13]. Since this tree is determined according to the value of the variable “ls”, we must verify that the number of jumps in this tree should not be arbitrarily long [11].

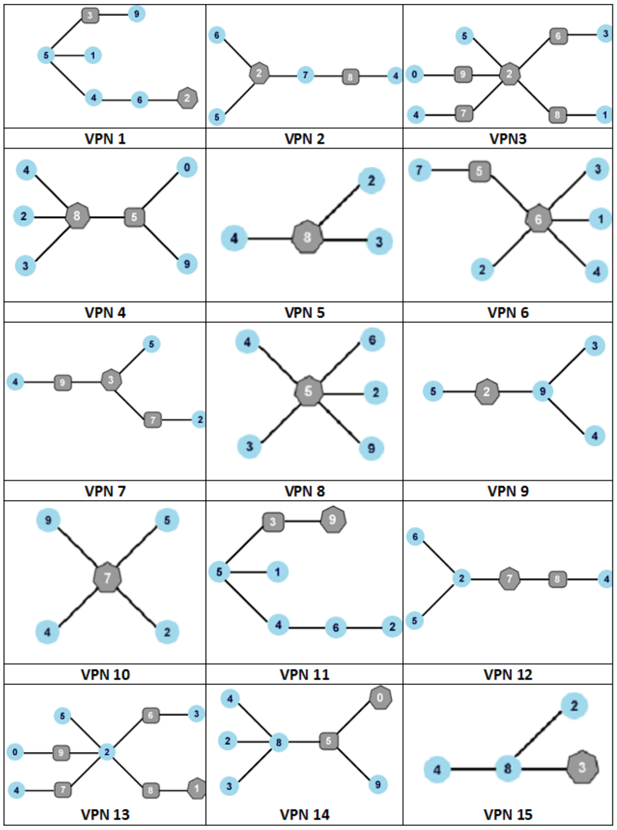


Fig. 1. Example of VPNs Generated by the TSR Algorithm.

IV. Simulations and Performance Analysis

This section presents the simulation input parameters for the different simulated VPN networks. All the simulation parameters are given in Table I. For accuracy and compliance, we ran each simulation scenario six times and averaged the measurements. Note that each of the six measurements conforms to the simulation parameters already described. To study the behavior of the two routing algorithms SPR and TSR according to the traffic intensity in the network, we varied the number of VPNs to be simulated. Fig. 1 gives an example of a VPN network that we have simulated. Each VPN is made up of a source and a set of destinations. Nodes in the form of a hexagon represent the sources. The nodes in the form of a circle represent the destinations. Rectangular nodes schematize transit nodes (Steiner) used to reach stations belonging to the same VPN.

We assume that data streams are sent from one source to a destination within the same VPN [14]-[21]. By applying the two heuristics SPR and TSR important differences in traffic distribution are remarkable. Both Fig. 2 and Fig. 3 represent an example of a scenario to be simulated where we have fixed the source and the destination while applying the two routing algorithms already described.

Fig. 2 schematizes a scenario where we have called the shortest path algorithm. By analyzing this scenario, we can see that the traffic going from source node 2 to destination node 4 is always focused on the same path, the shortest path (2-6 and 6-4).

TABLE I. SIMULATION PARAMETERS

Parameters	**Values**
Number of nodes	10
Link capacity	10 Mb/s
Delay transmission	10 ms
Number of source VPN	From 4 to 24 source nodes
Maximal size of window congestion	32
Simulation Time	400 seconds
Application type	FTP
Packet size	1000 bytes

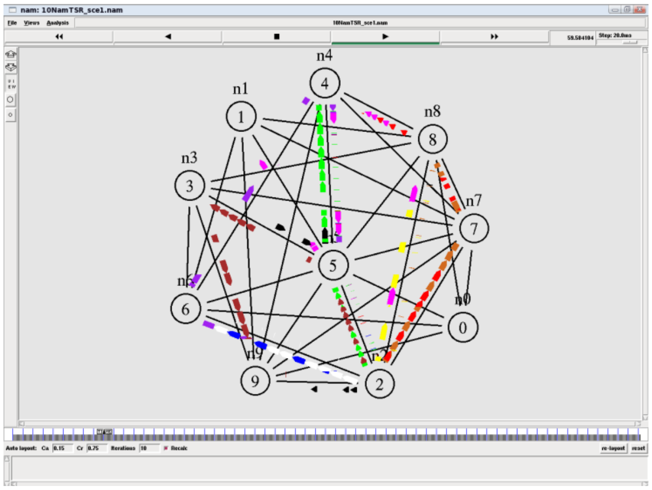


Fig. 2. SPR Traffic.

On the other hand, using the TSR heuristic, in Fig. 3, we notice that the traffic is shared in a more equitable way in the network. In fact, from source node 2 we can reach destination 4 by taking different paths (2-6 and 6-4 or even 2-5 and 5-4, etc…). This approach allows the maximum use of the network links. In order to be able to compare the two heuristics SPR and TSR in a more rigorous way, we will calculate the quality of service parameters: the average reception rate, the delay, the loss rate and the data flow sent.

Indeed, the routing technique will have a great influence on these parameters. This influence will be presented and highlighted later by the simulation results which concern the cases of 4 to 24 VPN sources representing respectively low and high traffic intensities.

A. *Average Reception Data Rate*

Fig. 4 details the flow variation for the two heuristics TSR and SPR. It clearly illustrates the speed changes depending on the number of VPN sources. Indeed, with 4 VPN sources, the heuristic SPR offers a throughput of 6.56Mbps while with TSR the throughput is 6.16Mbps. We can thus conclude that under a low traffic intensity, the application of the shortest path algorithm for traffic routing is more efficient than the application of traffic distribution.

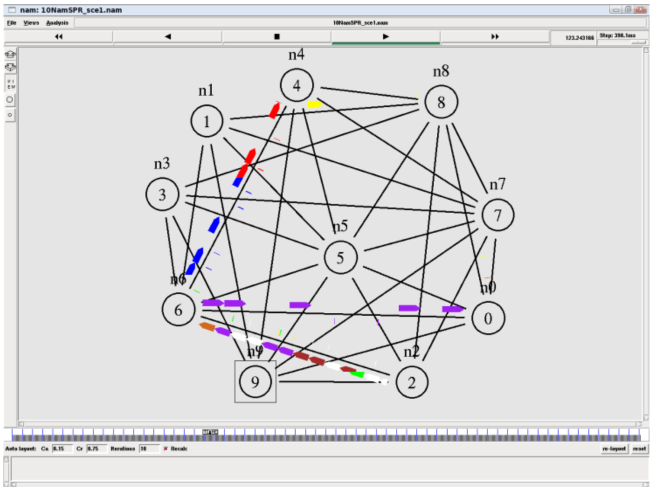


Fig. 3. TSR Traffic.

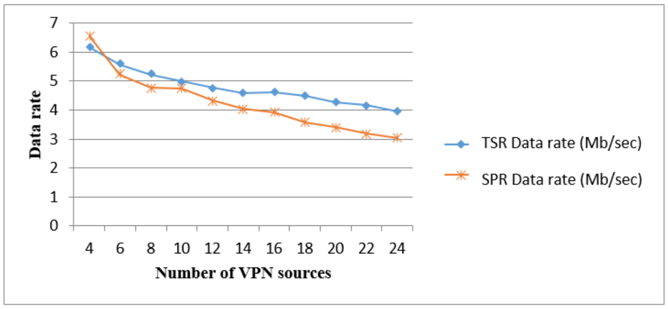


Fig. 4. Average Data Rate Reception.

Subsequently, with a high number of VPNs, the throughput with the SPR heuristic was significantly reduced to 4.74Mbps with 10 VPN traffic sources and 3Mbps with 24 VPN traffic sources. This decrease in throughput is due to the amount of traffic that is focused precisely on the shortest path.

On the other hand, with the TSR heuristic, we can notice that for 10 VPN sources the throughput is 5Mbps and for 24 VPN sources it can reach a value of 3.95Mbps. By comparing these results with the previous ones we deduce that the TSR algorithm offers a higher throughput especially for a large volume of traffic.

The analysis of the average reception rate allowed us to deduce that the TSR algorithm tends to use the maximum number of links, unlike the shortest path algorithm where the traffic always takes the shortest path which, in steady state, causes some links to become overloaded, leaving others unused. This had an influence on the flow.

Moreover, for a given throughput, the number of VPN sources admissible by the TSR method is significantly higher than that obtained with SPR. For example, if we want a speed of 4Mbps, with TSR we can admit up to 24 VPN sources while with SPR, this number is 14 sources.

B. Average End-To-End Delay

In this part we measure the average time taken for a 1000byte size packet to be transferred from a source to a destination. Fig. 5 shows the average delay for the two routing techniques used. This delay is given according to the number of VPN sources. Network reactions to the increased number of VPN sources for the two routing techniques are diverse. In fact, the average delay for low traffic intensity calculated with the shortest path algorithm exhibits a brief variation compared to that determined by the TSR algorithm. Indeed, for 4 VPN sources the average delay determined by the SPR approach is 24ms while with TSR this delay is 25ms.

By increasing the number of VPNs we can notice changes in the shape of the two curves. In fact, the average delay increases as a function of the number of VPN sources in an almost logarithmic fashion. However, these two curves look almost the same except that the average delay calculated by the TSR algorithm remains lower than that calculated by the SPR approach. Take for example the case of 16 VPNs where the average delay determined by the TSR heuristic is 27ms compared to that of SPR which is 37ms.

We define the gap as the difference between the average delay calculated for the same number of VPN sources for each of the two TSR and SPR heuristics. We notice that this gap is growing depending on the VPN sources (Fig. 5). From these results, we deduce that the difference between the delays obtained with the two approaches SPR and TSR is quite remarkable. This delay is reduced by applying the TSR heuristic because of the distribution of traffic over a large number of links, which offers more chances of going through small queues.

The fact of going through small queues means that the delay variation is smaller. With TSR, this is illustrated in Fig. 6. This figure gives a variation of delay for the two heuristics SPR and TSR. Indeed, as we have already mentioned during the throughput evaluation, the application of the SPR heuristic under a low traffic intensity is more efficient than that of traffic distribution. The curve presented in Fig. 6 confirms this result.

C. Loss Rate

We propose in the following to estimate the rate of lost packets for each routing technique. As shown in Fig. 7, a large gap between the loss rates obtained with the two heuristics TSR and SPR is perceived. Indeed, we can see that with low traffic intensity, the rate of packets lost by applying the shortest path algorithm is negligible. This rate increases as the number of VPN sources increases to reach a rate of 17×10^{-4} packets lost with 14 sources and 46×10^{-4} packets lost with 24 sources.

However, the traffic distribution algorithm has a higher loss rate than shortest path, which is 4×10^{-4} packets lost for 4 sources. Adding 10 more sources increased this loss rate to 16×10^{-4} lost packets. Similarly, we can notice that from this number of VPN sources (14 sources) the loss rate resulting from the use of the TSR heuristic becomes significantly lower than that obtained by the SPR algorithm.

From the results of the packet loss rate, we can conclude that the TSR algorithm gives lower loss rates for high load networks. Likewise, from Fig. 8, we see that the number of packets sent over the network using the distributed traffic routing technique is larger than that sent by applying the shortest path technique.

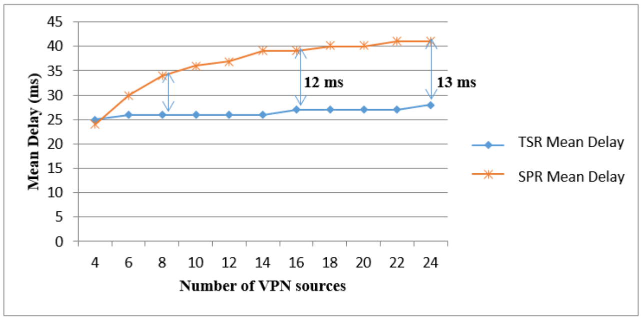


Fig. 5. Mean Delay.

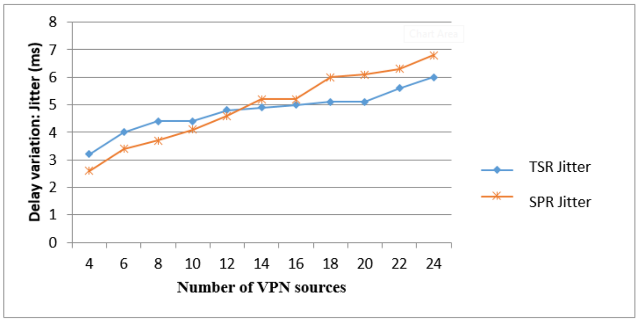


Fig. 6. Delay Variation.

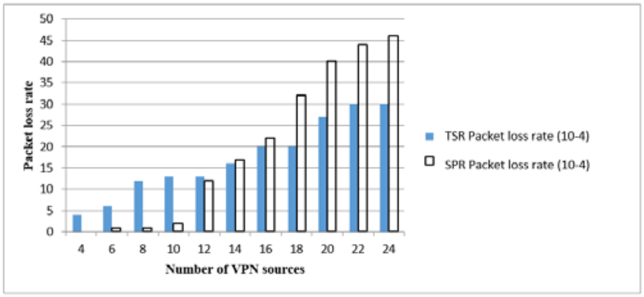


Fig. 7. Rate of Lost Packets.

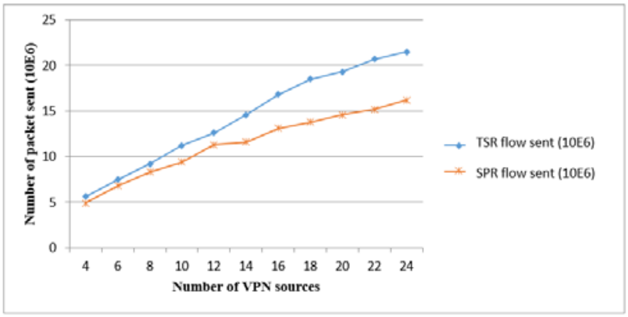


Fig. 8. Flow of Sent Data.

In fact, with the use of the shortest path algorithm, packets have a high chance of passing through overloaded queues, thus rejecting excess packets and therefore retransmission of rejected packets. Furthermore, with the use of the traffic distribution heuristic, there is a high probability of going through lightly loaded paths which results in a shorter routing time, a lower loss rate, as well as more packets sent.

V. Conclusion

This paper has been devoted to present a new routing approach, TSR (Traffic Split Routing) and to compare it to the classic routing of the shortest path SPR (Shortest Path Routing). In the first part, we presented different performance evaluation techniques. Then, we were interested in presenting different simulation tools and we justified our choice for the NS-2 tool. We went on to define the various quality of service parameters that we evaluated. Then, we detailed and analyzed the different simulation results obtained with different scenarios for the two approaches SPR and TSR.

Simulation results presented have demonstrated the effectiveness of TSR in the case of high traffic intensity. Thus, we were able to demonstrate that our approach, TSR, is more satisfactory for ensuring a better quality of service for certain types of applications such as real-time multimedia applications and VOIP which are very sensitive to the variation of speed and delay.

On the other hand, from the simulations, we noticed that the application of the traffic distribution algorithm allowed us to use a maximum of network resources. Indeed, for a scenario with 14 VPNs, we observed that the TSR heuristic used 69% of the network links. While with the shortest path heuristic, only 41% of all links in the network are used. In addition, the TSR approach makes it possible to accommodate a larger number of VPNs for a given objective (given loss rate, given throughput, etc.).

References

[1] B. WAXMAN. “Routing of Multipoint Connections”. IEEE J. on Selected Areas in Communications, numéro 9, volume 6, décembre, 1988, pages 1617-1622.

[2] Erdal Akin and Turgay Korkmaz. An Efficient Binary-Search Based Heuristic for Extended Unsplittable Flow Problem. In Computing, Networking and Communications (ICNC), 2017 International Conference on, pages 831–836. IEEE, 2017.

[3] Erdal Akin and Turgay Korkmaz. Routing Algorithm for Multiple Unsplittable Flows Between Two Cloud Sites with QoS Guarantees. In Computing, Networking and Communications (ICNC), 2017 International Conference on, pages 917–923. IEEE, 2017.

[4] Michael R Garey and David S Johnson. A Guide to The Theory of NP-Completeness. WH Freemann, New York, 70, 1979.

[5] Chen S. and Nahrstedt K., “An Overview of Quality of Service Routing for Next-Generation High Speed Networks: Problems and Solutions”. IEEE Network, 1998.

[6] Jaffe J. M., “Algorithms for Finding Paths with Multiple Constraints”. Networks, 1984. vol. 14: p. 95–116.

[7] J. W. Guck, A. Van Bemten, M. Reisslein, and W. Kellerer. Unicast QoS Routing Algorithms for SDN: A Comprehensive Survey and Performance Evaluation. IEEE Communications Surveys Tutorials, 20(1):388– 415, Firstquarter 2018.

[8] Verma S. Pankaj R., and leon-Garcia A., “QoS Based Multicast Routing for Multimedia Communications”. IEEE Workshop on QoS, 1997.

[9] Wu, Wenfei, Yoshio Turner, and Mike Schlansker. "Routing optimization for ensemble routing." 2011 ACM/IEEE Seventh Symposium on Architectures for Networking and Communications Systems. IEEE, 2011.

[10] J. Homer, S. Zhang, X. Ou, D. Schmidt, Y. Du, S. R. Rajagopalan, and A. Singhal. Aggregating vulnerability metrics in enterprise networks using attack graphs. Journal of Computer Security, 21(4):561–597, 2013.

[11] X. Ou, W. F. Boyer, and M. A. McQueen. A scalable approach to attack graph generation. In Proceedings of the 13th ACM conference on Computer and communications security, pages 336–345. ACM, 2006.

[12] LeMay, E., Scarfone, K., Mell, P.: The common misuse scoring system (CMSS): Metrics for software feature misuse vulnerabilities. US Department of Commerce, National Institute of Standards and Technology (2012).

[13] Ou, X., Singhal, A.: Security risk analysis of enterprise networks using attack graphs. In: Quantitative Security Risk Assessment of Enterprise Networks, pp. 13–23. Springer (2011).

[14] P. S. Sarker, V. Venkataramanan, D. S. Cardenas, A. Srivastava, A. Hahn and B. Miller, "Cyber-Physical Security and Resiliency Analysis Testbed for Critical Microgrids with IEEE 2030.5," 2020 8th Workshop on Modeling and Simulation of Cyber-Physical Energy Systems, Sydney, Australia, 2020, pp. 1-6, doi: 10.1109/MSCPES49613.2020.9133689.

[15] M. Bagaa, D. L. C. Dutra, T. Taleb and K. Samdanis, "On SDN-Driven Network Optimization and QoS Aware Routing Using Multiple Paths," in IEEE Transactions on Wireless Communications, vol. 19, no. 7, pp. 4700-4714, July 2020, doi: 10.1109/TWC.2020.2986408.

[16] Faycal Bensalah and Najib El Kamoun, “Novel Software-Defined Network Approach of Flexible Network Adaptive for VPN MPLS Traffic Engineering” International Journal of Advanced Computer Science and Applications(IJACSA), 10(4), 2019.

[17] R. A. Mishra, A. Kalla, K. Shukla, A. Nag and M. Liyanage, "B-VNF: Blockchain-enhanced Architecture for VNF Orchestration in MEC-5G Networks," 2020 IEEE 3rd 5G World Forum (5GWF), Bangalore, India, 2020, pp. 229-234.

[18] Kohei Arai, “Routing Protocol based on Floyd-Warshall Algorithm Allowing Maximization of Throughput” International Journal of Advanced Computer Science and Applications(IJACSA), 11(6), 2020.

[19] J Wognin Vangah, Sié Ouattara, Gbélé Ouattara and Alain Clement, “Global and Local Characterization of Rock Classification by Gabor and DCT Filters with a Color Texture Descriptor” International Journal of Advanced Computer Science and Applications(IJACSA), 10(4), 2019.

[20] V. Q. Rodriguez, F. Guillemin and A. Boubendir, "5G E2E Network Slicing Management with ONAP," 2020 23rd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), Paris, France, 2020, pp. 87-94.

[21] S. Yucel, "Algorithmic Framework for QoS and TE in Virtual SDN Services," 2019 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2019, pp. 1494-1499.

A Platform for Extracting Driver Behavior from Vehicle Sensor Big Data

Sultan Ibrahim bin Ibrahim[1], Emad Felemban[2]
Ahmad Muaz Qamar[4], Abdulrahman A. Majrashi[6]
Computer Engineering Department
College of Computing and Information Systems
Umm Al Qura University, Saudi Arabia

Faizan Ur Rehman[3], Akhlaq Ahmad[5]
Institute of Consultation and Research Studies
Umm Al-Qura University, Makkah
Saudi Arabia

***Abstract*—Traffic analysis of vehicles in densely populated areas and places of public gathering can provide interesting insights into crowd behavior. Hajj is a spatio-temporally bound religious activity that is held annually and attended by more than 2 million people. More than 17,000 buses are used to transport pilgrims on fixed days to fixed locations. This poses great challenges in terms of crowd management. Using Global Positioning System (GPS) and Automatic Vehicle Location (AVL) sensors attached to buses, a large amount of spatio-temporal vehicle data can be collected for traffic analysis. In this paper, we present a study whereby driver behavior was extracted from an analysis of vehicle big data. We have explained in detail how we collected data, cleaned it, moved it to a big data repository, processed it and extracted information that helped us characterize driver behavior according to our definition of aggressiveness. We have used data from 17,000 buses that has been collected during Hajj 2018.**

Keywords—GPS Data; AVL sensors; hajj; big data; traffic analysis

I. Introduction

Open source traffic data for Saudi Arabia, provided by the Saudi Ministry of Interior, shows that rash driving is a major factor contributing to road accidents [1] in the Kingdom of Saudi Arabia [1]. The data also shows that the maximum number of incidents happen in the Makkah region, whose population is far less than the most populous Riyadh region. One of the possible reasons for this increase of accidents in the Makkah region is the fact that the annual Hajj pilgrimage happens in this area whereby more than 2 million people visit the region from all corners of the world. A large fleet of vehicles is needed to transport these people between cities and the holy places. To understand the behavior of drivers, we used the data of 17,000 buses collected in Hajj 2018.

Hajj is an annual pilgrimage of Muslims that happens every year from the 8th to the 13th of Dhul-Hijjah, the 12th month [2][3] of the lunar Islamic calendar. More than 2 million people from Saudi Arabia and around the world come to perform Hajj. International pilgrims come a few days earlier to the city of Makkah. On the 8th, they leave for Mina (bounded with a red dashed line in Fig. 1), a small dwelling of permanently installed tents near Makkah. On the morning of the 9th, they move to Arafat (bounded with a blue dashed line in Fig. 1), an open space about 17 km further south. After sunset, they come back and spend the night at Muzdalifah (bounded with a green dashed line in Fig. 1), completing their return trip to Mina by the morning of the 10th. From 10th onwards, for a period of 3 days, including an optional 4th one, the pilgrims stone the three pillars called Jamarat. They also go to slaughterhouses and visit the Grand Mosque in Makkah to perform certain obligatory rituals [2]. All this movement is restricted with respect to time and space. A fleet of more than 17,000 buses is required to move the pilgrims across the holy places, collectively referred to as Mashaer. To collect traffic data from the large number of buses utilized by the pilgrims, the General Syndicate of Cars (Naqaba[2]) has ordered the bus operators to attach Automatic Vehicle Location (AVL) sensors to their buses. Data collected from these AVL sensors has proven to be very useful in studying a number of characteristics related to traffic. We have developed a system that utilizes this data to present interactive visualization to perceive traffic activity at various hours of the day throughout the Hajj season.

This article provides an expansion of our previous research [5][6] by focusing on analysis of sensor data to extract driver behavior. In this paper we present the results of a study conducted on pilgrim bus data. The main objective of this study is to understand drivers' aggressive behavior by capturing data stamped with spatial and temporal information using Automatic Vehicle Location (AVL) devices based on Global Positioning System (GPS). The vehicles used in this study were pilgrim buses were equipped with AVL sensors that provided location-based data. From this data, we extracted speeds of various buses on different routes. We used this speed data as well as other parameters extracted from the source data to classify drivers and their driving skills according to our defined driver profiles.

This paper is divided into six sections. Section 2 discusses the state of the art in the area of traffic data collection for information extraction using AVL and GPS technology. Section 3 explains the methodology for collecting data and extracting useful information. Section 4 presents an overview of the system architecture and explains the role of different components in the system. Section 5 discusses the results of

[1] https://data.gov.sa/Data/en/organization/ministry_of_interior_-_general_directorate_of_traffic

[2] https://www.haj.gov.sa/en/InternalPages/Details/92

applying analytics on data. Section 6 concludes the paper with a summary of the whole study.

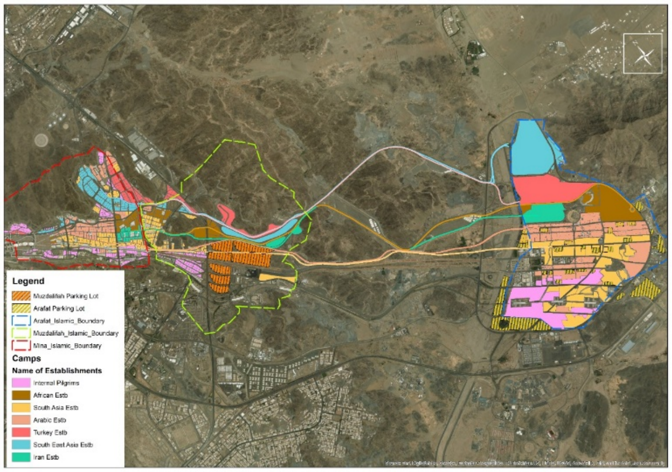


Fig. 1. Map of Mina (Red Boundary), Muzdalif (Green Boundary) and Arafat (Blue Boundary) near Makkah City (Buildings in the Lower Left).

II. Literature Review

Our study entails using GPS data collected from AVL devices to detect driver behavior from the obtained data. We detail below the state of the art in this regard.

Grengs et al. [7] explained the procedural challenges to collect, store and design databases and to manipulate and analyze the enormous set of geocoded data captured from trips and tours for understanding driving characteristics of a single driver for a duration of a month or so. They studied 78 drivers by using an automobile and recorded their behavior on a day-to-day basis for about a month. They added the position coordinates and time stamp with each data set. Their results showed that the travel patterns were more complex when compared to traditional travels.

Necula et al. [8] utilizes a Hidden Markov Model (HMM) method and a training process and presented an interactive tool to study drivers' behaviors. The tool integrates the past real data captured from various local drivers and analyzes the routes followed by every driver utilizing time, distance height and speed information. The tool also manages the maximum likelihood to validate the next route segment for a network of roads.

Feng et al. [9] examined the merits of using accelerometer with GPS data in a transportation study. They presented three approaches by first considering accelerometer only, then GPS sensors only and lastly a combination of both accelerometer and GPS sensors. They utilized Bayesian Belief Network model to study the three different transportation modes and found that the use of accelerometer can successfully play a significant role in imputing transportation mode. The single usage of each device separately was helpful in terms of predictivity accuracy. The combined usage of both the devices reviled best performance.

Choi et al. [10] considered real driving scenarios and presented a model to detect distraction due to peripheral tasks. They utilized Hidden Markov Models (HMMs) and captured drivers' characteristics using CAN-Bus (Controller Area Network) sensor. This provided them a variety of information such as steering wheel angles, brake status and brake usage with respect to time as well as breaking behavior with associated speed, etc. They defined the drivers' behavior in terms of action and distraction based on the abovementioned data.

Warwick et al. [11] presented their study on drowsiness of drivers and quoted about 100,000 crashes a year on national highways. They emphasized the development of smart a system to detect the drowsiness earlier to avoid accidents. They compared the causes of accidents because of vehicle-based issues with drivers' physiological-based approaches and found that causes were mostly due to drivers' physiology. They proposed to design a driver drowsiness detection system which utilizes wireless wearables to measure a driver's physiological data. The sensory setup provides data that can be analyzed to find key parameters related to drivers' drowsiness and generates early alerts to act in time.

Jasinski et al. [12] proposed a method that identifies real-time aggressive behavior of drivers and generate prior alerts for any dangerous behavior which may result in a severe accident. The method composed of four stages - data collection, pre-processing, semantic enrichment and calculations to compute the aggressiveness of drivers. The (TAI - Trajectory Aggressively Indicator) aggressive Indicator values varies from 0 to 100, with 0 means no aggressiveness and 100 the extremely aggressive. The proposed approach also considers the environmental conditions in calculating a better estimate of TAI.

Paefgen et al. [13] developed a method to measure the accident risk. The method utilizes GPS data collected from a large sample of traffic data from a telematics provider in northern Italy, where there were 1500 drivers with and without accidents over a period of two years. The GPS trajectories were analyzed to study the driver risk profiling problem and their findings in this regard were promising.

Khan at al. [14] presented a comprehensive survey on driving activities, the reasons for accidents, and systems to generate prior notification for drivers for their safety and comfortable drive upon early detection of an accident. Based on their findings, they suggested that a well-designed DMAS (driving monitoring and assistance system) can improve critical issues associated with drivers as well as the challenges associated with the related driving environment.

Stutts et al. [15] and Kan et al. [16] described the main reasons for majority (about 90%) of road accidents as, distraction, fatigue and aggressive driving style. Distraction refers to eating or drinking, looking at off road people, getting busy in small activities like sharing food, texting or attending phone calls, etc. and causes more than half of the accidents. Fatigue explains the physical condition of drivers like drowsiness, over acting to show up extra driving skills, etc. Aggressiveness is related to driving style, overreaction to overtaking cars, or applying breaks in front of other vehicles. Their study suggested the use of DMAS by considering the factors associated with drivers and the driving environment.

Jingqiu et al. [17], developed a hybrid model to study driving behavior and risk patterns. The model utilizes Autoencoder and Self-organized Maps (AESOM) approach to extract driving behaviors. They made 4032 observations by collecting data through GPS sensors, in Shenzhen, China, and analyzed the speed and excessive acceleration and summarize their findings as that AESOM usage may improve the quality of the driving.

Arumugam et al. [18] presented a comprehensive survey on driving behavior and addressed drivers' agressive behavior and detailed multiple incidents on short and long-term driving activities. The purpose of the survey was to explore the solution to minimize the risk on roads by considering drivers' emotional factors defining their driving behavior and provide the information to insurance companies to profile drivers' behavior and define the best possible insurance premium package for risk prevention.

Improving transportation system is one of the significant requirements for large gatherings where crowd safety is major concern [4]. For this researcher proposed several intelligent transportation systems, which in turn opened the door for research areas related to traffic data collection, data mining [19][20][21] processing and analysis [22]. GPS (global position system) sensors are valuable data sources which help in tracking the vehicles by reading their spatial information in real time [23].

III. Methodology

We have divided our process into 5 major categories - data management, computation, behavior definition, comparison and interpretation. Each category has been further divided as shown in Fig. 2. Data management includes data collection, data cleaning, data enrichment and data visualization. Computation consists calculation of distance, speed and acceleration. Definition entails defining speed ranges with respect to roads. Comparison of the bus speed with allowed ranges with respect to road. Interpretation includes identifying other parameters, driver characteristics and classifying drivers.

A. *Data Management*

Data management is a critical step that directly helps in improving the quality of extracted information for analytics. It includes cleaning, structuring, removing noise or fixing missing data and validation. In the following, we explain how we performed the above operations on our data.

1) Data collection: To develop any system for decision-making requires collection of a good amount of historical data. Our data source (Naqaba) is the transport authority of the Ministry of Hajj and Umrah that collected data of 17,000+ buses using automatic vehicle location (AVL) service providers in Hajj 2018. Fig. 3 shows some facts about the collected data. The data collected was for pilgrim movement on different routes during Hajj, such as, Jeddah Airport-Makkah, Makkah-Madinah, Madinah – Madinah Airport, Makkah – Mina, Makkah – Arafat, Arafat – Muzdalifah, Muzdalifah – Mina.

2) Data cleaning: We excluded noise from the data using the spatial boundary algorithm that removes locations outside a given boundary from the dataset. We also removed data entries where difference in distance between two points for the same bus is around zero and the location is not on main roads, assuming the bus to be parked at a pickup or drop off point or at a parking location.

3) Data enrichment: In our case, we collected the GPS traces of the buses during Hajj. Generally, AVL sensor providers configure GPS devices to transmit data at intervals varying from 2 to 7 minutes in length. We found that sometimes the duration between the two locations from the same bus is up to 20-30 minutes due to which some entries are lost, mostly because of connectivity issues. Fig. 4 shows the anomaly associated with a missing entry or a long distance between recorded points. The two locations from the same bus are shown in Fig. 4. The black line shows the line as per the raw data and the blue line shows an extra data point after data enrichment, i.e., after adding a missing point. The enriched data shows the actual distance and will be beneficial to extract knowledge for analytics.

4) Data validation: Along with data enrichment, we have also performed data quality checks to handle the GPS error issues. An error of only 10 meters can show the location of the bus on the other side of the road that will lead to a large error in calculating distance, and henceforth speed and acceleration.

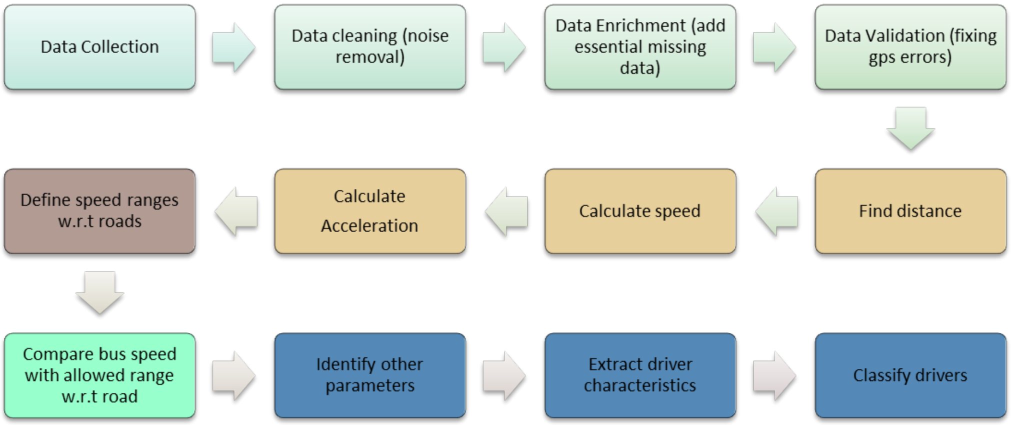

Data Collection
Data cleaning (noise removal)
Data Enrichment (add essential missing data)
Data Validation (fixing gps errors)
Find distance
Calculate speed
Calculate Acceleration
Define speed ranges w.r.t roads
Compare bus speed with allowed range w.r.t road
Identify other parameters
Extract driver characteristics
Classify drivers

Fig. 2. Overview of the Methodology.

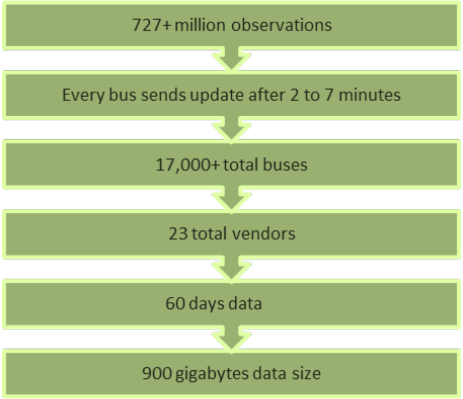


Fig. 3. Overview of Collected Data.

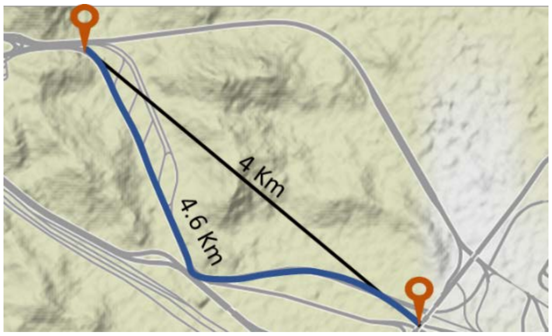


Fig. 4. Enriched Data in Blue Colour.

B. Computation

After data completing manipulation, we enriched the dataset by adding acceleration and distances information as additional data columns, using mathematical equations as follows:

Calculate acceleration (a):

$$a = \frac{\Delta v}{\Delta t} = \frac{(vf - vi)}{(tf - ti)} \qquad \frac{m}{s^2}$$

Where vf = final velocity, m/s; vi = initial velocity, m⁄s; Δv = difference in velocity, m⁄s; tf = ending time in seconds, ti = starting time in seconds, and Δt = difference of time in seconds.

Calculate distance (d) between adjacent points on the globe as shown in Fig. 5 by using Haversine formula as shown in Fig. 5:

$$a = \sin^2\left(\Phi B - \Phi\frac{A}{2}\right) + \cos \Phi A * \cos \Phi B * \sin^2\left(\lambda B - \frac{\lambda A}{2}\right)$$

$$c = 2 * \operatorname{atan2}\left(\sqrt{a}, \sqrt{(1-a)}\right)$$

$$d = R \cdot c$$

Where Φ = latitude, λ = longitude, R = radius of the earth (R ≈ 6.371 km), A= ending point, B = initial point and d = the distance between two points.

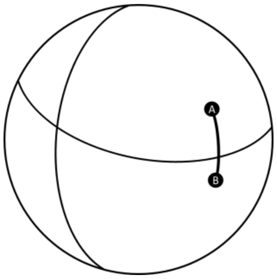


Fig. 5. Distance between Two Points on the Globe is Calculated using the Haversine Formula Due to the Curvature of the Earth's Surface.

TABLE I. ROAD TYPES

Road	Type
Highway	Motorway
Highway	Trunk
Highway	Primary
Highway	Secondary
Highway	Residential
Highway	Unclassified

C. Definition

We use the open source Open Street Maps (OSM) and extract road related characteristics for each road such as the speed limit to calculate the speed threshold for each road and highway as shown in Table I. The speed limit varies from 60-140 kmph depending on the road type and its proximity to populated areas. We allow the driver to cross the speed limit up to 10%.

D. Comparison

Based on the street profiles extracted above, we match the bus speed data with our speed threshold for each road on the route to classify the vehicles according to speed. The spatial queries have been used with the help of a spatial relational database. The spatial relational database stores the geometry of each road and spatial query checks whether the location of the bus belongs to that road segment or not.

E. Interpretation

After separating the vehicles violating traffic rules from others, we apply the spatio-temporal conditions mentioned previously on data to classify the drivers into aggressive and non-aggressive behavior.

IV. System Architecture

Fig. 6 shows the high-level view of the big data platform that we developed to analyze the bus data. We developed a data lake layer that consist a Master data service in addition to an MS SQL service. The master data service provides a visualization of all the relations in the data based on different parameters, such as Establishment, Offices or bus number (every bus is assigned to an Office which is under an Establishment related to a geographical area). The MS SQL database contains the original data we received from Naqaba.

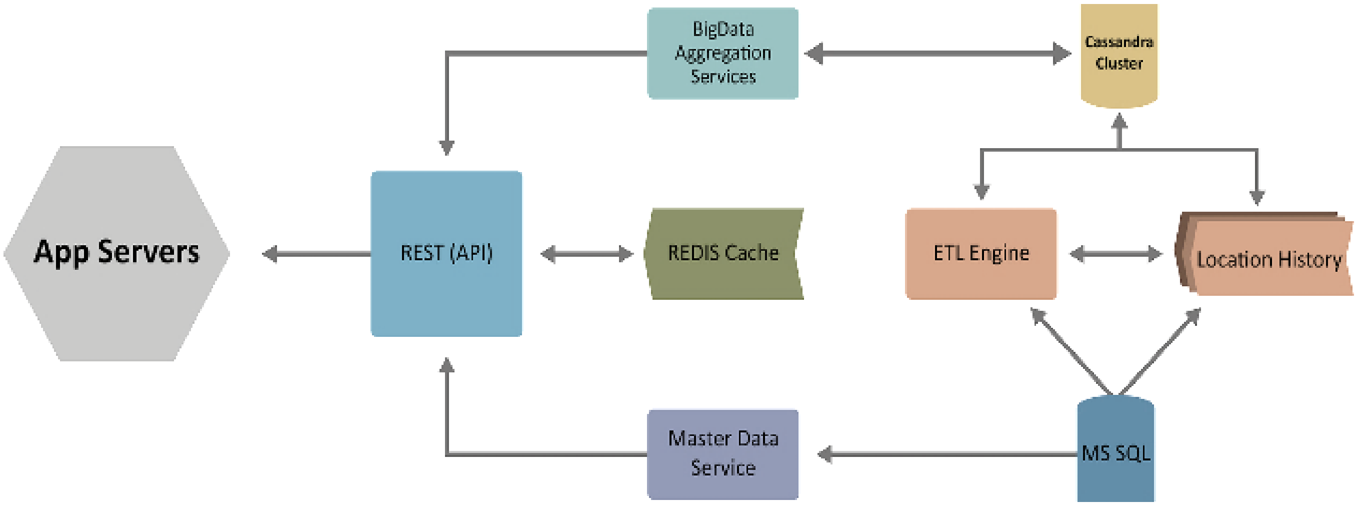


Fig. 6. Overview of the Big Data Platform.

Our big data layer is made up of a Cassandra cluster and a Big Data Aggregation service. We have migrated Location History data to the Cassandra cluster for cleaning and removing noise using the ETL engine. The benefit of using Cassandra cluster is that it increases efficiency and scalability using a distributed, wide column store, running on a NoSQL database management system.

We have used Hadoop and Presto to setup the big data aggregation service. Presto is efficient tool used for distributed SQL analytical queries on data in the Hadoop distributed file system (HDFS). Hadoop is highly beneficial for batch-based analytics while Cassandra is good for time-based. *REDIS* cache is an open-source (BSD-licensed), in-memory data structure store used as a cache.

It is good for caching a huge number of key-value pairs. We have use REDIS cache to boost the performance of the system. We have devised a RESTful API that provides a list of APIs to handle requests coming from the front-end. The front-end requests the API to fetch data from the Master data service, the big data aggregation service or from the REDIS cache and returns the results that are visualised on the screen.

The API Server provides the front-end data visualization and analysing service. It allows the user to display data based on multiple filters, including Establishment (Mo’assasa) name, Office (Maktab) number, company name, bus number and route.

Fig. 7 compares time taken to perform queries in MS SQL server and our big data platform. The orange line shows the time to fetch the records of bus id in MS SQL server while the grey line shows the time to fetch the records in our big data platform. It is clear from the Fig. 8 that time gained from moving to the big data platform is significant.

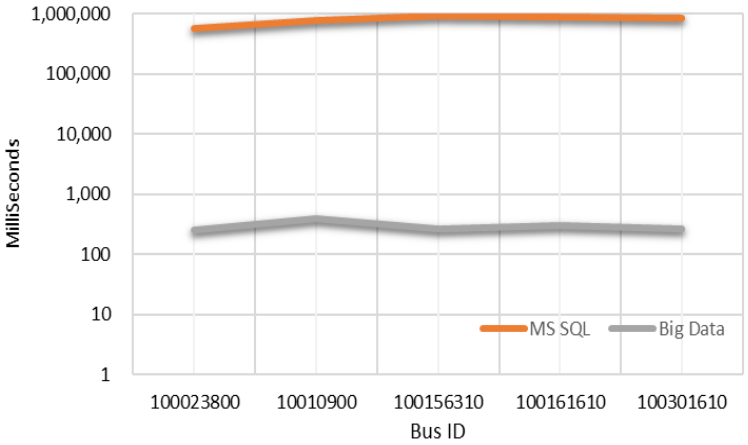
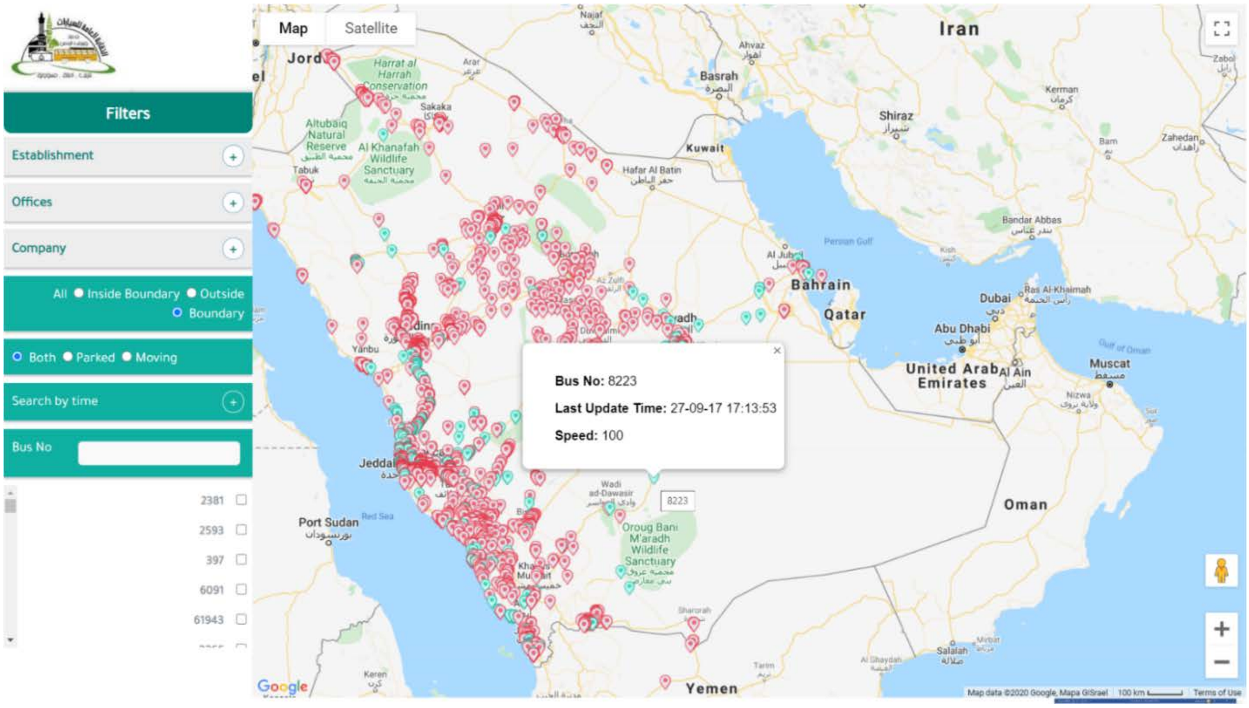


Fig. 7. Time Comparison before and after Migration to Big Data Platform.



Bus No: 8223

Last Update Time: 27-09-17 17:13:53

Speed: 100

Fig. 8. Platform Visualization after Migration to Big Data Framework.

V. Analytics and Discussion

Each bus was allotted to a single driver for the entire Hajj season. The vehicles were tracked by capturing their spatio-temporal information and the collected data was analyzed by considering each road speed limit. Table II is a snapshot of the collected data along with few violations' information.

First, we selected the continuous number of observations that exceeded speed limit threshold (80kmh) with starting & ending timestamp. Then we calculated the duration of violation from starting and ending timestamp in minutes and seconds as described in the above table. Fig. 9, is the summary of the number of violations detected by the AVL sensors. We can observe that for most of the days the driver's behavior was aggressive, crossing the threshold speed several times.

Further, we have classified the violations based on severity, and collected information regarding the frequency, duration and severity of speed limit violation by a driver. Table III details our classification of violations.

Fig. 10 summarizes the recorded violations of a driver with perspective of above classification during the entire Hajj season. We can see that normal category of violations is common, which shows that 22 times the driver violated the speed limit but just for a few seconds or so, and then reduced his speed less than the threshold. To address normal violation cases, it happened 8 times that he violated the speed limit for about 10 minutes duration. Three times he continuously violated for a duration between 10-20 minutes and twice, he committed severe violations, that is, for more than 20 minutes.

TABLE II. Violation Information Extracted from Source Data

Start Timestamp	End Timestamp	No of Consecutive Observations	Violation Duration in (s)	Time Minutes
2016-09-13 04:47:44	2016-09-13 04:55:44	2	480.0	8
2016-09-13 06:05:46	2016-09-13 06:13:45	2	479.0	7.98
2016-09-13 06:31:55	2016-09-13 06:31:55	1	0.0	0
2016-09-14 05:51:27	2016-09-14 05:51:27	1	0.0	0
2016-09-15 06:29:29	2016-09-15 06:37:29	2	480.0	8
2016-09-15 11:43:35	2016-09-15 12:19:36	17	2161.0	36.01
2016-09-15 12:29:43	2016-09-15 12:35:37	3	354.0	5.90

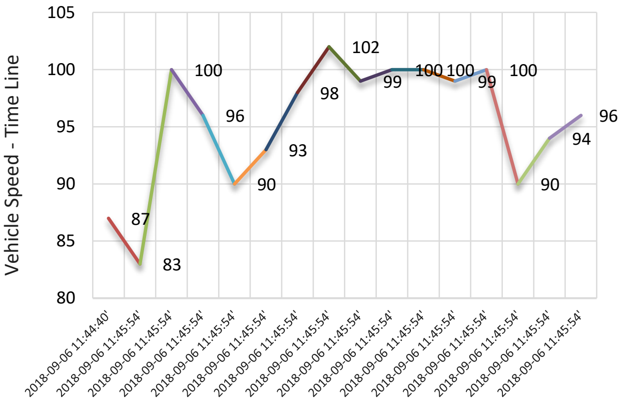


Fig. 9. Vehicle Speed - Time Line.

TABLE III. VIOLATION SEVERITY

Severity	Explanation
Once	A violation just for a few seconds and not a continuous one for a long duration. These violations occurred only in individual observations so we cannot get starting timestamp & ending timestamp, hence we called it Once
Normal	If the violation's duration is less than 10 minutes, then we called it Normal violation
High	If the violation's duration is between 10 to 20 minutes, then we say it is High
Severe	If the violation's duration is more than 20 minutes, then we consider it Severe

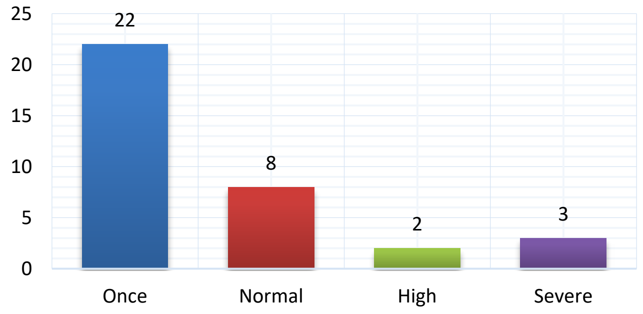


Fig. 10. Driver's behavior (Violation Classification-Bus No. 1).

On the same scale, the system analyzed mobility of several buses and found that the bus with ID 152, violated the speed limit in total 905 times, out of which mostly the violation was in once category and 197 times the bus violated for duration less than 10 minutes. No case of severe violation was recorded for bus 152. This analysis helps find out the worst cases as shown in Fig. 11.

Among all the drivers, there were some with best performance as they committed no or minor violations. Fig. 12 below is the summary of the best buses where drivers' behavior was in a satisfactory range. The best case is for bus ID who violated just for a few seconds during the entire Hajj season.

Fig. 13 shows the speed-based detection of aggressiveness. The geographical locations captured with timestamps were analyzed for one of the drivers, who was driving on C-ring road, in Makkah. Upon analysis, we discussed both the non-aggressive behavior (Fig. 13a: the green dots show that the driver's speed was within the threshold value and was not committing any violation) and aggressive behavior (Fig. 13b: the red dots show that the driver violated several times the threshold speed showing his aggressive behavior). The data points show that the driver's behavior was aggressive for 52.55% of the collected data points, which has been visualized in the figure below. The corresponding heading and acceleration information for both the cases is also detailed.

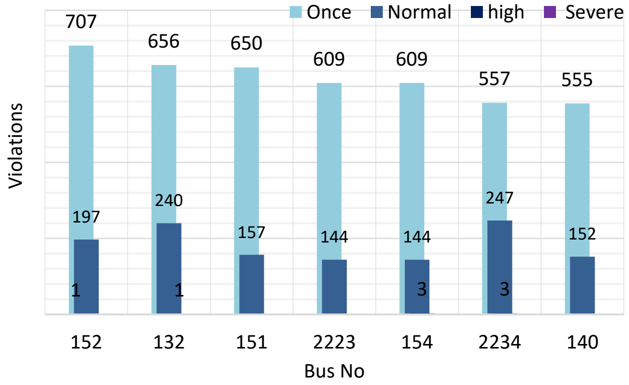


Fig. 11. Violation Severity based on Bus Number. Identification of Worst Drivers.

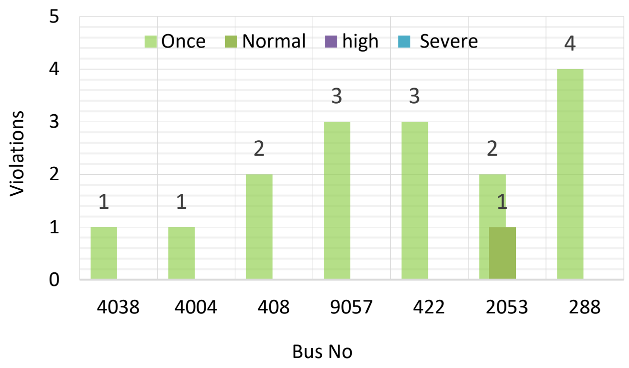
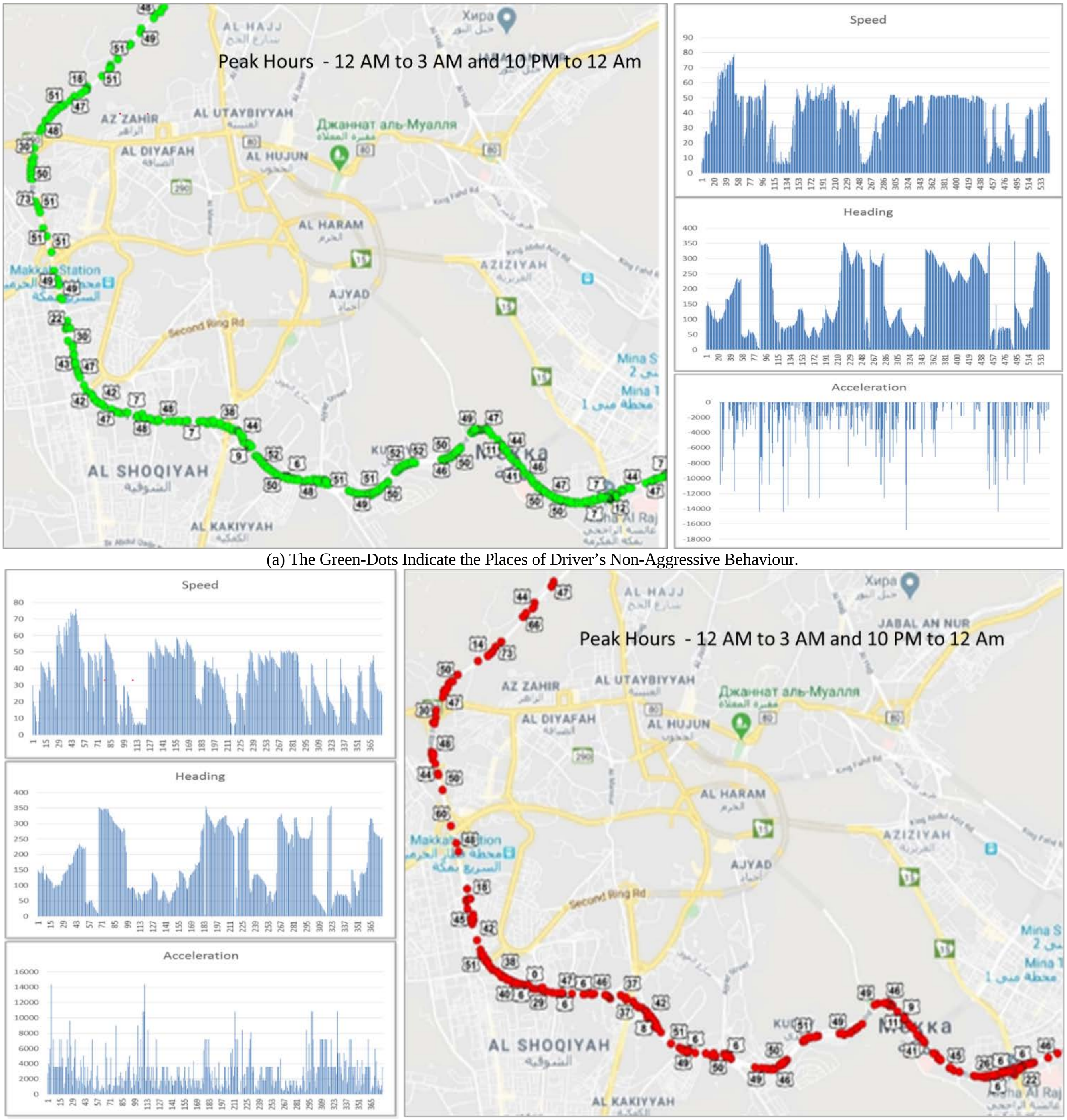


Fig. 12. Identification of Best Driver.



(a) The Green-Dots Indicate the Places of Driver's Non-Aggressive Behaviour.

Fig. 13. (b): The Red-Dots Indicate the Places of Driver's Aggressive Behaviour.

VI. Conclusion

In this paper we have presented a study to extract driver aggressiveness information from GPS data obtained through AVL sensors attached to a fleet of 17,000+ buses used to transport pilgrims from one place to another in the holy areas during the Hajj pilgrimage. We have explained the details of data preparation and pre-processing methodology we have adopted by moving the data to a big data platform for efficient query processing. We have also highlighted our procedure for extraction of information. One of the limitations of the experiment is that it was carried out in one particular city on drivers from outside the city. However, our technique is generic and can be used on any AVL based data in any part of the world.

Acknowledgement

The authors would like to thank Deanship of Scientific Research and Prince Khalid Al-Faisal Chair for Developing Makkah Al-Mukarramah and the Holy Places at Umm Al-Qura University (project # DSRUQU.PKC-41-5) for the financial support. We would also like to thank Eng. Osama Fateh from Naqaba for providing us with the bus data.

References

[1] U. B. Ghaffar and S. Ahmed, “A Review of Road traffic accident in Saudi Arabia: the neglected epidemic,” Indian J. Forensic Community Med., vol. 2, no. 4, p. 242, 2015, doi: 10.5958/2394-6776.2015.00010.7.

[2] A. Ahmad, M. A. Rahman, M. Ridza Wahiddin, F. Ur Rehman, A. Khelil, and A. Lbath, “Context-aware services based on spatio-temporal zoning and crowdsourcing,” Behav. Inf. Technol., 2018, doi: 10.1080/0144929X.2018.1476586.

[3] A. Ahmad et al., “A framework for crowd-sourced data collection and context-aware services in Hajj and Umrah,” in 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA), Nov. 2014, pp. 405–412, doi: 10.1109/AICCSA.2014.7073227.

[4] F. Ur Rehman et al., "A constraint-aware optimized path recommender in a crowdsourced environment," 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), Marrakech, 2015, pp. 1-8, doi: 10.1109/AICCSA.2015.7507185.

[5] E. Felemban, F. U. Rehman, A. A. Biabani, A. Naseer, and U. AlAbdulwahab, “Towards Building an Interactive Platform for Analyzing Movement of Buses in Hajj,” in 2019 IEEE International Conference on Big Data (Big Data), Dec. 2019, pp. 3775–3778, doi: 10.1109/BigData47090.2019.9005521.

[6] E. Felemban, F. U. Rehman, H. Wadood, and A. Naseer, “Towards Building Evacuation Planning Platform using Multimodal Transportation for a Large Crowd,” in 2019 IEEE International Conference on Big Data (Big Data), Dec. 2019, pp. 4063–4066, doi: 10.1109/BigData47090.2019.9006226.

[7] J. Grengs, X. Wang, and L. Kostyniuk, “Using GPS Data to Understand Driving Behavior,” J. Urban Technol., vol. 15, no. 2, pp. 33–53, Aug. 2008, doi: 10.1080/10630730802401942.

[8] E. Necula, “Mining GPS Data to Learn Driver’s Route Patterns,” in 2014 16th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, Sep. 2014, pp. 264–271, doi: 10.1109/SYNASC.2014.43.

[9] T. Feng and H. J. P. Timmermans, “Transportation mode recognition using GPS and accelerometer data,” Transp. Res. Part C Emerg. Technol., vol. 37, pp. 118–130, Dec. 2013, doi: 10.1016/j.trc.2013.09.014.

[10] S. Choi, J. Kim, D. Kwak, P. Angkititrakul, and J. H. L. Hansen, “Analysis and classification of driver behavior using in-vehicle can-bus information,” Bienn. Work. DSP In-Vehicle Mob. Syst., no. October 2015, pp. 17–19, 2007.

[11] B. Warwick, N. Symons, X. Chen, and K. Xiong, “Detecting Driver Drowsiness Using Wireless Wearables,” in 2015 IEEE 12th International Conference on Mobile Ad Hoc and Sensor Systems, Oct. 2015, pp. 585–588, doi: 10.1109/MASS.2015.22.

[12] M. G. Jasinski and F. Baldo, “A Method to Identify Aggressive Driver Behaviour Based on Enriched GPS Data Analysis,” GEOProcessing 2017 Ninth Int. Conf. Adv. Geogr. Inf. Syst. Appl. Serv., no. March 2017, pp. 97–102, 2017, [Online]. Available: https://www.thinkmind.org/download.php?articleid=geoprocessing_2017_6_20_38002.

[13] J. Paefgen, F. Michahelles, and T. Staake, “GPS trajectory feature extraction for driver risk profiling,” TDMA’11 - Proc. 2011 Int. Work. Trajectory Data Min. Anal., pp. 53–56, 2011, doi: 10.1145/2030080.2030091.

[14] M. Q. Khan and S. Lee, “A Comprehensive Survey of Driving Monitoring and Assistance Systems,” Sensors, vol. 19, no. 11, p. 2574, Jun. 2019, doi: 10.3390/s19112574.

[15] J. C. Stutts, D. W. Reinfurt, and E. A. Rodgman, “The role of driver distraction in crashes: an analysis of 1995-1999 Crashworthiness Data System Data.,” Annu. Proc. Assoc. Adv. Automot. Med., vol. 45, no. May, pp. 287–301, 2001.

[16] M. Q. Khan and S. Lee, “A comprehensive survey of driving monitoring and assistance systems,” Sensors (Switzerland), vol. 19, no. 11, 2019, doi: 10.3390/s19112574.

[17] J. Guo, Y. Liu, L. Zhang, and Y. Wang, “Driving Behaviour Style Study with a Hybrid Deep Learning Framework Based on GPS Data,” Sustainability, vol. 10, no. 7, p. 2351, Jul. 2018, doi: 10.3390/su10072351.

[18] S. Arumugam and R. Bhargavi, “A survey on driving behavior analysis in usage based insurance using big data,” J. Big Data, vol. 6, no. 1, p. 86, Dec. 2019, doi: 10.1186/s40537-019-0249-5.

[19] S. Turner and L. Albert, “Its Data Quality Control and the calculation of Mobility Performance Measures,” 2000.

[20] Chaogui Zhang, Zhiyong Zheng, Fuqiang Zhang, and Jiangtao Ren, “Multidimensional traffic GPS data quality analysis using data cube model,” in Proceedings 2011 International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE), Dec. 2011, pp. 307–310, doi: 10.1109/TMEE.2011.6199204.

[21] G. Gecchele, R. Rossi, M. Gastaldi, and A. Caprini, “Data Mining Methods for Traffic Monitoring Data Analysis: A case study,” Procedia - Soc. Behav. Sci., vol. 20, pp. 455–464, 2011, doi: 10.1016/j.sbspro.2011.08.052.

[22] H. Wang, M. Ouyang, Q. Meng, and Q. Kong, “A traffic data collection and analysis method based on wireless sensor network,” Eurasip J. Wirel. Commun. Netw., vol. 2020, no. 1, 2020, doi: 10.1186/s13638-019-1628-5.

[23] L. Shen and P. R. Stopher, “Review of GPS Travel Survey and GPS Data-Processing Methods,” Transp. Rev., vol. 34, no. 3, pp. 316–334, May 2014, doi: 10.1080/01441647.2014.903530.

APPENDIX 1: SAMPLE OF ENRICHMENT DATA

Door_NO	Bus_No	GPS_Date	Latitude	Longitude	GPS Speed	Heading	Acceleration	Status	Distance	Calculated Speed (Km/h)	Compare speeds	Speed Accuracy
2345	1006	02-SEP-16 05.34.48.671000000 PM	21.439383	39.542163	10	67	126.0190919	Aggressive	0.53333	8	Calculated_Speed < GPS Speed	80.00%
2345	1006	02-SEP-16 05.34.49.461000000 PM	21.439386	39.542193	14	90	9856.262834	Aggressive	0.00568	14	Calculated_Speed = GPS Speed	100.00%
2345	1006	02-SEP-16 05.34.50.486000000 PM	21.439373	39.54223	16	115	4845.222073	Aggressive	0.00619	15	Calculated_Speed < GPS Speed	93.75%
2345	1006	02-SEP-16 05.34.51.498000000 PM	21.439343	39.542263	17	137	2403.204272	Aggressive	0.00791	19	Calculated_Speed > GPS Speed	89.47%
2345	1006	02-SEP-16 05.34.52.204000000 PM	21.439301	39.54229	18	148	2990.033223	Aggressive	0.00602	18	Calculated_Speed = GPS Speed	100.00%
2345	1006	02-SEP-16 05.34.53.428000000 PM	21.439253	39.542315	21	154	7563.02521	Aggressive	0.00873	22	Calculated_Speed > GPS Speed	95.45%
2345	1006	02-SEP-16 05.35.00.599000000 PM	21.438873	39.542495	22	151	473.7465456	Aggressive	0.35	21	Calculated_Speed < GPS Speed	95.45%
2345	1006	02-SEP-16 05.35.01.501000000 PM	21.438823	39.542528	23	144	2398.401066	Aggressive	0.01001	24	Calculated_Speed > GPS Speed	95.83%
2345	1006	02-SEP-16 05.35.02.475000000 PM	21.438775	39.54257	25	135	4881.355932	Aggressive	0.01024	25	Calculated_Speed = GPS Speed	100.00%
2345	1006	02-SEP-16 05.35.03.364000000 PM	21.438731	39.54262	24	128	-2639.296188	Non-Aggressive	0.00834	22	Calculated_Speed < GPS Speed	91.67%
2345	1006	02-SEP-16 05.35.04.221000000 PM	21.438691	39.542675	27	123	8845.208845	Aggressive	0.00848	25	Calculated_Speed < GPS Speed	92.59%
2345	1006	02-SEP-16 05.35.06.698000000 PM	21.438603	39.542796	29	129	2668.64344	Aggressive	0.02098	28	Calculated_Speed < GPS Speed	96.55%
2345	1006	02-SEP-16 05.35.08.348000000 PM	21.438503	39.542911	29	133	0	Non-Aggressive	0.01957	30	Calculated_Speed > GPS Speed	96.67%
2345	1006	02-SEP-16 05.35.09.323000000 PM	21.438445	39.542966	31	137	5442.176871	Aggressive	0.01103	30	Calculated_Speed < GPS Speed	96.77%
2345	1006	02-SEP-16 05.35.10.404000000 PM	21.438386	39.543021	30	140	-2564.102564	Non-Aggressive	0.01131	29	Calculated_Speed < GPS Speed	96.67%
2345	1006	02-SEP-16 05.35.12.651000000 PM	21.438268	39.543116	27	146	-4073.934364	Non-Aggressive	0.02136	29	Calculated_Speed > GPS Speed	93.10%
2345	1006	02-SEP-16 05.35.14.719000000 PM	21.438158	39.543195	24	149	-3972.048547	Non-Aggressive	0.01813	24	Calculated_Speed = GPS Speed	100.00%
2345	1006	02-SEP-16 05.35.15.449000000 PM	21.438105	39.543226	24	152	0	Non-Aggressive	0.00886	22	Calculated_Speed < GPS Speed	91.67%
2345	1006	02-SEP-16 05.35.16.349000000 PM	21.438051	39.543255	22	156	-5337.286879	Non-Aggressive	0.00899	24	Calculated_Speed > GPS Speed	91.67%

APPENDIX 2: SAMPLE OF FINAL RESULTS

Bus_No	Data Collection Time	Data Calculation Time	Total time	Number of Records	Aggressive	Non-Aggressive
1006	0.0:0.0:0.030506	0.0:0.0:0.000005	0.0:0.0:0.030511	5658	42.08%	57.92%
1010	0.0:0.0:0.039205	0.0:0.0:0.000004	0.0:0.0:0.039210	7346	38.32%	61.68%
1029	0.0:0.0:0.023313	0.0:0.0:0.000004	0.0:0.0:0.023317	4368	40.38%	59.62%
1102	0.0:0.0:0.008098	0.0:0.0:0.000003	0.0:0.0:0.008101	1652	14.41%	85.59%
1109	0.0:0.0:0.029801	0.0:0.0:0.000004	0.0:0.0:0.029805	5575	43.34%	56.66%
1202	0.0:0.0:0.000013	0.0:0.0:0.000002	0.0:0.0:0.000015	3	0.00%	100.00%
1205	0.0:0.0:0.012375	0.0:0.0:0.000003	0.0:0.0:0.012378	2311	40.07%	59.93%
1207	0.0:0.0:0.058354	0.0:0.0:0.000005	0.0:0.0:0.058359	10858	43.67%	56.33%
1209	0.0:0.0:0.015614	0.0:0.0:0.000005	0.0:0.0:0.015619	2916	40.98%	59.02%
1210	0.0:0.0:0.058767	0.0:0.0:0.000005	0.0:0.0:0.058772	11019	34.42%	65.58%
1211	0.0:0.0:0.047687	0.0:0.0:0.000005	0.0:0.0:0.047691	8966	36.17%	63.83%
1213	0.0:0.0:0.026305	0.0:0.0:0.000004	0.0:0.0:0.026308	5010	36.71%	63.29%
1214	0.0:0.0:0.026867	0.0:0.0:0.000005	0.0:0.0:0.026872	5155	33.40%	66.60%
1215	0.0:0.0:0.000034	0.0:0.0:0.000002	0.0:0.0:0.000036	8	25.00%	75.00%
1216	0.0:0.0:0.018076	0.0:0.0:0.000003	0.0:0.0:0.018079	3386	40.84%	59.16%
1217	0.0:0.0:0.054429	0.0:0.0:0.000005	0.0:0.0:0.054434	10162	38.98%	61.02%
1219	0.0:0.0:0.025904	0.0:0.0:0.000005	0.0:0.0:0.025909	4848	36.14%	63.86%
1220	0.0:0.0:0.040444	0.0:0.0:0.000004	0.0:0.0:0.040448	7569	38.79%	61.21%
1222	0.0:0.0:0.013775	0.0:0.0:0.000003	0.0:0.0:0.013777	2610	39.46%	60.54%
1223	0.0:0.0:0.030309	0.0:0.0:0.000005	0.0:0.0:0.030313	5654	38.80%	61.20%
1225	0.0:0.0:0.011557	0.0:0.0:0.000004	0.0:0.0:0.011561	2162	38.34%	61.66%
1226	0.0:0.0:0.004240	0.0:0.0:0.000003	0.0:0.0:0.004243	817	33.90%	66.10%
1228	0.0:0.0:0.039108	0.0:0.0:0.000003	0.0:0.0:0.039111	7338	41.69%	58.31%
1229	0.0:0.0:0.058615	0.0:0.0:0.000005	0.0:0.0:0.058620	10974	36.66%	63.34%

Simulation and Analysis of Variable Antenna Designs for Effective Stroke Detection

Amor Smida*
Department of Medical Equipment Technology, College of Applied Medical Sciences
Majmaah University, AlMajmaah 11952, Saudi Arabia
Microwave Electronics Research Laboratory, Department of Physics Faculty of Mathematical
Physical and Natural Sciences of Tunis, Tunis ElManar University, Tunis 2092, Tunisia

***Abstract*—The variety of applications of patch antenna for portable applications has opened the avenues for the possibilities of having compact, cost-efficient, and life-saving devices. Considering the challenges of portability and cost in making it feasible for detecting strokes in the masses of developing countries where the demand is quite high, this study builds the groundwork for such device fabrication. In total five antenna designs were investigated for their assessment in identifying the stroke. Two main studies of electromagnetic wave interaction and bio-heating of the human head phantom had been accomplished and the results are compared. The main comparison and identification of the stroke location with the human head phantom are presented by the specific absorption rate (SAR), both visualized as volumetric plot and stacked contour slices for clarifying the shape and positioning of the stroke in vertical and horizontal dimensions. The results show that the SAR values for Antenna A & D are the lowest with the values of 1.44 x 10-5 W/kg and 1.96 x 10-5 W/kg, respectively. But the induced electric field and isothermal temperature achieved were highest by Antenna D, with values of 0.25 emw and 133.92 x 10-8 K, respectively; and, the 2-D far-field radiation patterns confirmed better performance by it amongst all others. Hence, making the Antenna D as the most preferred choice for the prototyping stage. The overall trade-off of key parameters is studied herein in this simulation study and based on that the most suitable antenna design is proposed for the experimental prototype testing. The results suggest that the simulation results give a clear insight into the feasibility of stroke detection with the proposed setup and presents high viability for portable, low-cost, and rapid stroke detection applications.**

Keywords*—Stroke detection; Specific Absorption Rate (SAR); Patch Antenna; S-parameter (S11); electromagnetic wave; bio-heat transfer

I. Introduction

In the past decade, the WHO declared cerebrovascular accidents (stroke) as the first reason for neurological dysfunction in the western world and based on worldwide statistics second most critical reason for mortality and the third rising reason for disability [1]. The sudden disruption of blood flow in some parts of the brain occurs due to a blockage or rupture that causes the death of brain cells and in the severe case may even lead to dementia and acute depression [2]. Every year globally almost 5 million people die and the other 5 million are rendered disable due to such stroke-related accidents. Though the major impact is seen on the low and middle-income group countries, wherein 70% of the total stroke cases are reported and 87% of the total deaths and disability-adjusted life years (DALYs viz. a term combining years spent with disability and years of life lost) occurs [3-5]. The disparity in the number of reported cases between the developing and developed countries is almost seven times [3].

Currently, the guidelines for stroke management are to undergo a series of treatments as per the resulting diagnosis. But there are relatively only two widely used methods available for stroke diagnosis i.e. X-ray based computer tomography (CT) and magnetic resonance imaging (MRI) [6]. Both are only available in the most advanced health centers and involve a huge cost of setup establishment, hence very scarcely available for the ongoing demand of diagnosis. Of which the MRI is even rarer due to its overall make complexity, cost and likewise the challenging operable skills demanded [7]. Even CT scans demand a greater skill for efficiently interpreting the results by a practicing professional, consequently creating delays or misinterpretations. And the adverse effects of X-ray make it even harder for continuous monitoring due to the limited safety exposure limits of these ionizing radiations [8]. If strokes can be identified at a very early stage, then accompanied by suitable remedial measures and prevention strategies the rate of stroke mortality can be effectively reduced [9, 10]. Moreover, the demand for effective, portable, and cost-effective system is the main research question currently that can address all the above-listed limitations away.

Considering the emerging demands stroke monitoring and detection have become a focus for developing more cost-effective and easily implementable techniques. Of which the Microwave Imaging (MWI) seems very promising due to its ease of setup and operation; with the capability to give a quick diagnosis as a complementary tool for the prehospital record on the type, size, shape, in some cases the number of blood blockages and as a harmless setup for continuous monitoring of the effect due to ongoing treatment since no harmful irradiations are involved herein [11-14]. MWI takes the concept of variable reflections from surfaces with different permittivity and conductivity and helps build a 3D mapping of the model under study. This concept is now explicitly applied for mapping the human head phantom with the idea to locate the stroke at early stages with clear visualization of its shape, size, and most importantly their numbers with each one's

*Corresponding Author

precise location. Many prototype devices are now been proposed and developed based on this concept [15-25]. Some are now even popularly applied as a commercial solution [18, 23, 24].

Since the effectiveness of the MWI is mainly dictated by the number and type of antennas employed involved the efforts for portable device development are mostly focused on tuning the involved parameters. In the lieu of getting improved results, there have been cases wherein almost 177 antenna arrays were tested for successful head model reconstruction and stroke identification [18]. Such expansive use of antennas imposes a serious challenge on the practical implementation of the model whilst prototyping and commercializing the concept. Hence, the need for a much powerful but lesser number of antenna involvement. The study herein proposes a concept for analyzing the human head phantom for stroke identification using an array of only 8 patch antennas, making it more viable for practical implementation. Though there are five different antenna designs studied for understanding the suitability of a design for its effectiveness in stroke identification. The simulation study presented herein compares the performance of each antenna design and gives clear practical reasoning for selecting one of them for future experimental studies. Also, the results clearly present the concept for generating 3D results for identifying the stroke shape, size, and location effectively. The combined analysis of the chosen antenna designs along with the 3D model concept for stroke identification makes this study different from the existing ones [16, 17, 20, 22, 25] and paves the way for streamlined laboratory testing with the best performing design.

In this paper, the concept of stroke identification and its location is presented along with an analysis of the antenna design most suitable to use. Among the five antennas it designed for 2.4 GHz band and lower SAR values and making them a favorable choice. Further, the antenna designs are also simulated to observe the effect on the reflection parameter/gain and the changes of SAR value with different designs.

II. Numerical Setup

A. Antenna Designs

For this study variable antenna designs were chosen for comparing their performances in terms of suitability for stroke identification, though considering the future manufacturing constraints not very intricate models were used. As it would be easy to make the intricate layouts and test them in simulation, but the next prototyping stage might face various issues and eventual delays due to limitations in manufacturing capability; in Fig. 1 The different antenna layout designs layouts with all the key dimensions of all the antenna designs studied. These were modeled separately in ProE CAD software to provide greater flexibility in designing the various dimensional aspects.

B. Simulation Setup

The simulation is setup within the COMSOL Multiphysics 5.4 version software. The imported head phantom geometry used for this simulation study is recreated following the reference specific anthropomorphic mannequin (SAM) as mentioned in IEEE 1528 [26] and IEC 62209-1 [27]. It was developed by the IEEE Standards Coordinating Committee 34, Subcommittee 2, Working Group 1 (SCC34/SC2/WG1). It was chosen for this study since it is the widely used model for radiation related study on the human head and is followed by renowned organizations like the Association of Radio Industries and Businesses in Japan [28], European Committee for Electrotechnical Standardization (CENELEC) [29], the US Federal Communications Commission [30], etc. Though to reduce the problem size the model is scaled down by 60% with minor adjustments, to reduce the mesh complexity and save computational cost. Using the properties of cortical bone tissue within an ellipsoid geometry the simplified brain model was rendered to embody the human head phantom.

Further, each antenna design as described above was imported in COMSOL and was defined with materials as a layer of metal for the patch and FR4 dielectric material for the baseboard including ground. The metal patch was defined as a perfect electric conductor (PEC) having a negligible loss, within the radio frequency (RF) module’s electromagnetic waves frequency domain interface. And the power source is represented by a lumped port fed by 50 Ω boundary condition. 8 x antennas in array form were arranged around the phantom at a radius of 130 mm. The stroke was modeled as a spherical (10 mm radius) mass of blood with the appropriate permittivity and conductivity values. To avoid unwanted reflection, the head phantom containing the stroke along with the antenna array were all enclosed within a perfectly matched layer (PML) spherical air domain. Which absorbs all outgoing waves, acting as an anechoic chamber and represented that the testing was done within infinite open space.

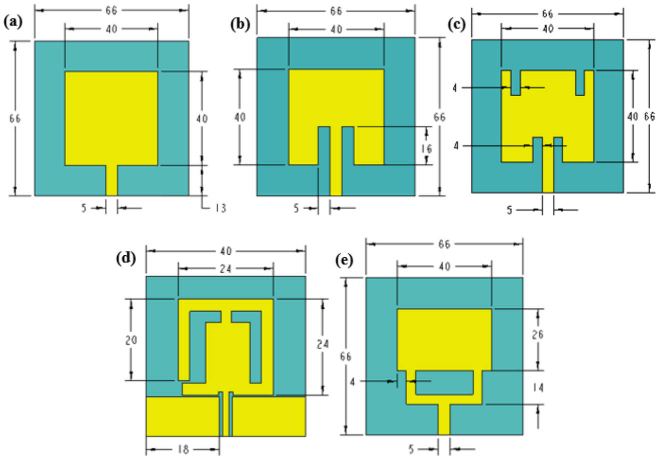


Fig. 1. The different Antenna Layout Designs Tested in this Simulation Study for the Human Head Stroke Detection by Potable Setup Application.

The analysis was carried out as two studies viz. bio-heating and electromagnetic (EM) wave interaction of the human head phantom. The volumetric interpolation model sampling function takes data from an MRI scan processed as an image data file with 109 slices and finally imported as a text file [31]. This text file relates the variation of tissue type within the created phantom model as to the real human head. And the exact material properties were taken from reference [32] which are summarized in Section 1 (S1) of Supporting Information. The patch antenna excited by lumped port emits the radiation which is directed onto the phantom by the array arrangement. The working frequency was chosen as 2.45 GHz as the general norm for on body and in body applications stated by industrial, scientific, and medical (ISM) bandwidth is in the range of 2.4 – 2.5 GHz [33].

Temperature distribution was solved by Penni's bioheat transfer equation, which mathematically models the physical heating phenomenon of the living tissues. This is the widely used model for heat interaction studies of human tissues [34-38] and is given by the following Equation 1.

$$\nabla \cdot (k\nabla T) + q_p + q_m - \omega c_b (T - T_a) = \rho c_p \frac{\partial T}{\partial t} \qquad (1)$$

where k is the thermal conductivity of tissue (W/m/°C); T and T_a is the local tissue temperature and arterial temperature in °C respectively; q_p and q_m is the energy deposition and metabolism rate (W/m^3) respectively; ω is the local blood perfusion rate (kg/m^3/s) of tissue; c_b and c_p are the specific heat (J/kg/°C) of blood and tissue material respectively; ρ is the density (kg/m^3) of tissue material; and t is the time (sec).

The EM wave interaction was solved by the vector-Helmholtz equation at the selected frequency, which is give as below by Equation 2.

$$\nabla \times \frac{1}{\mu_r} \nabla \times E - k_0^2 \varepsilon_r E = 0 \qquad (2)$$

where E represents the field, μ_r is the relative permeability, k_0 is the wave vector in free-space and ε_r represents the permittivity.

Finally, the specific absorption rate (SAR) is calculated which represents the RF energy absorbed by tissue in per unit time and is calculated from the tissue density and the electromagnetic dissipation density. This is important to estimate the usability of such setup from the safety point of view, which is done by comparing the maximum exposure values with the prescribed limits. It is estimated by applying the following Equation 3.

$$\mathrm{SAR} = \sigma \frac{|E|^2}{\rho} \qquad (3)$$

where σ represents the tissues electric conductivity, $|E|$ is the electric field (RMS) norm, and ρ is the tissue density. The SAR value is calculated herein in watts per kilogram (W/kg).

The finite element method (FEM) method is used to solve the governing partial differential equations (PDEs) which is formulated in COMSOL Multiphysics. The energy exchange between the metabolism and blood perfusion is been overlooked for future ex-vivo testing. The patch and phantom model were meshed with free tetrahedral form elements having 68574 degrees of freedom, while the remaining boundary regions were built with free triangular mesh.

III. Results and Discussions

The average relative permittivity plot is more like a clarification that the model is properly generated and that the comparison models are properly correlated with the results. The mean relative permittivity (emw) contour graphs for all antenna designs are shown in Fig. 2 and were recreated using imported MRI image data [31] from which permittivity values are determined. The stroke is also seen here, showing the cross section of the stroke by taking the required reference plane.

The bioheat transfer study converts the incident waves into heat energy which is solved in the chosen COMSOL Multiphysics RF module. Based on the solutions derived from this analysis the temperature distribution plots are generated as shown by Fig. 3 (surface temperature plot which tells only the surface effect as the name suggests) and Fig. 4 (sub-surface temperature plot, which gives a detailed understanding of the induced heating). The noted surface temperature rise (°C) values due to the incident EM waves induced heating is 8.95 x 10^{-7}, 4.45 x 10^{-7}, 5.02 x 10^{-7}, 14 x 10^{-7}, and 5.41 x 10^{-7} for when the antenna-A, antenna-B, antenna-C, antenna-D, and antenna-E respectively are used. This implies that the antenna-D is creating more surface heating effects amongst all the designs. Considering the iso-surface heating effect due to the incident waves, the temperature rise (°C) values are 85.02 x 10^{-8}, 42.34 x 10^{-8}, 48.80 x 10^{-8}, 133.92 x 10^{-8}, and 51.39 x 10^{-8} for when the antenna-A, antenna-B, antenna-C, antenna-D, and antenna-E respectively are used. Again, here also the antenna-D stands out as the most heating type design. Though considering the absolute values of the temperature rise, the rise is still not very high to create an adverse effect.

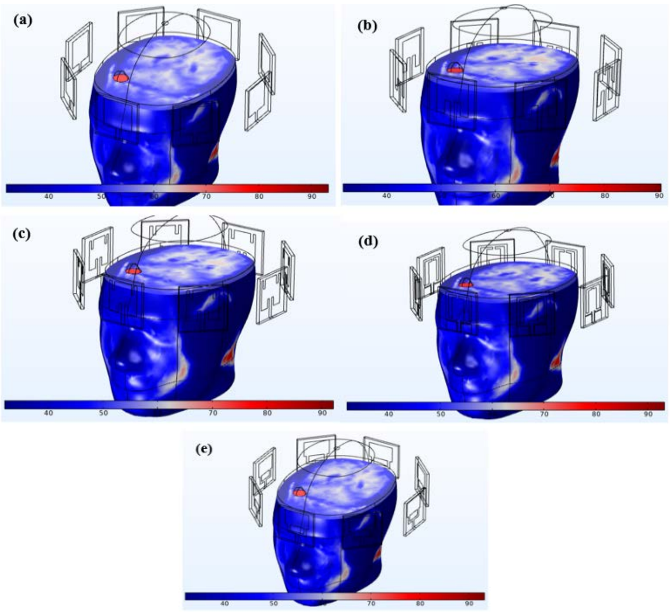


Fig. 2. The Average Relative Permittivity (emw) Contour Plots of the Stroke Containing Human Head Phantom under the Influence of Microwave Irradiation by Various Antenna Designs viz. (a) Antenna-A, (b) Antenna-B, (c) Antenna-C, (d) Antenna-D, and (e) Antenna-E.

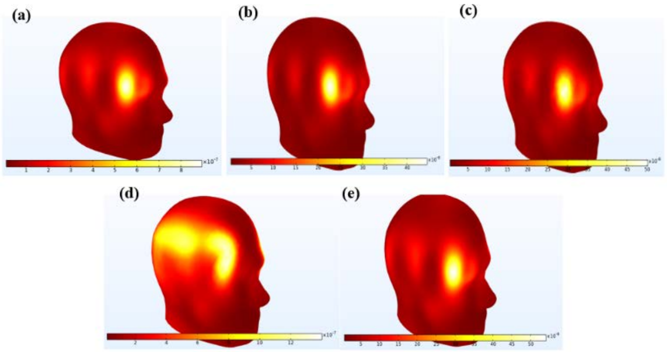


Fig. 3. Surface Temperature Plots of the Stroke Containing Human Head Phantom under the Influence of Microwave Irradiation by Various Antenna Designs viz. (a) Antenna-A, (b) Antenna-B, (c) Antenna-C, (d) Antenna-D, and (e) Antenna-E.

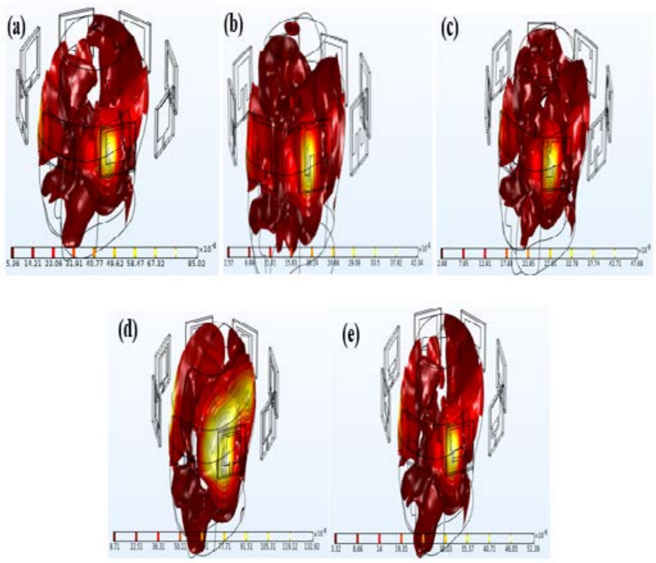


Fig. 4. Isothermal (Isosurface) Contours of the Stroke Containing Human Head Phantom under the Influence of Microwave Irradiation by Various Antenna Designs viz. (a) Antenna-A, (b) Antenna-B, (c) Antenna-C, (d) Antenna-D, and (e) Antenna-E.

The following results present the SAR rate by the head phantom due to the induced waves from the antenna array. The highest SAR value is observed near the regions of the phantom that are directly facing or nearby the antenna's incident electric field. Usually, the amount of SAR recorded depends on the number of antenna and their positioning & dielectric properties. The regions of the head phantom have distinct values of dielectric properties (permittivity and conductivity) which varies with the induced frequency and geometry of the antenna. Hence, the SAR which varies for the RF dependent human tissues; depends on the change in antenna design when the frequency is kept constant. This functionality is exploited for the key aspect of stroke identification since the blood clotted region responds differently than the surrounding tissue mass.

The simulated SAR fields on the human head phantom having the stroke, for the five antenna designs arranged in circumferential manner are shown in Fig. 5 (volumetric plot) and Fig. 6 (slices contour). By analysing each SAR slice individually, the exact shape and location of the stroke can be concluded. For example, the layered SAR slices (front view) arrangement is included in Section S2 in Supporting Information. The observation was carried out on the whole head phantom at 2.45 GHz. This arrangement of observation gives a unified way of locating the stroke position, both horizontally and vertically. This also benefits with the shape identification of the stroke for variable sizes, shapes, and positions. Higher SAR value hence is not desirable because it creates uneasiness and adverse health effects [39], also its increased value hints at higher resistive losses [40].

From all the result plots of the SAR exposure in W/g as shown in Fig. 7 the SAR value is higher near the antenna premises (near field area), where existence of strong electric field is obvious. The maximum SAR values (in W/kg) observed are 1.44x10^{-5}, 7.56 x 10^{-6}, 8.12 x 10^{-6}, 1.96 x 10^{-5}, and 8.63 x 10^{-6} for when the antenna-A, antenna-B, antenna-C, antenna-D, and antenna-E respectively are used. All the observed values are quite lower than the described safety limit of 1.6 – 2 W/g for 1g of human tissue model [41].

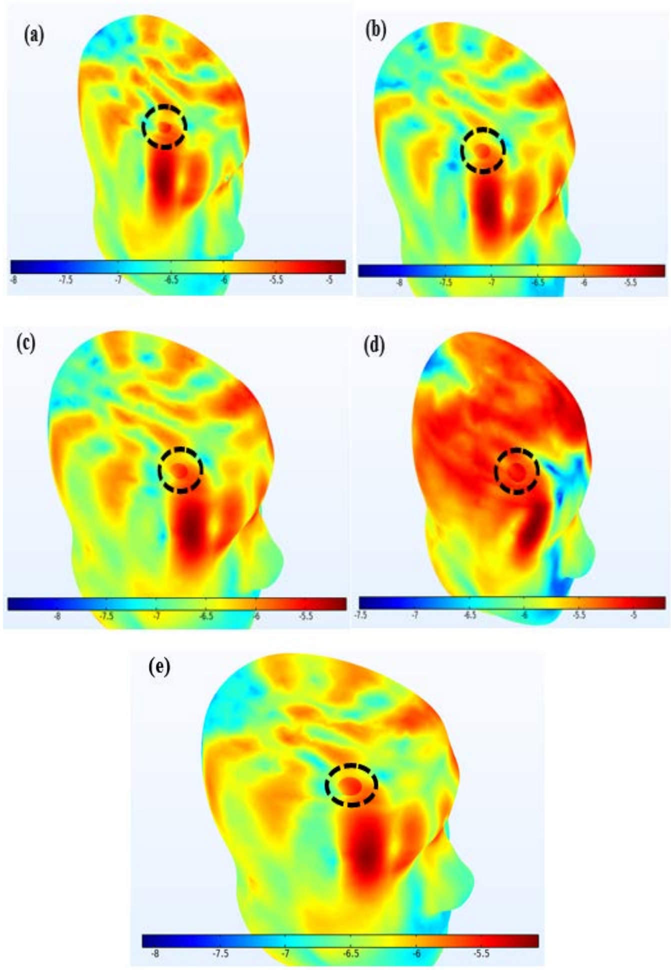


Fig. 5. Horizontal Cross-Sectional Layout of the SAR (Volumetric) Log-Scale Plot Showing the Exact Stroke Location in the Top view (the Exact Positioning of the Stroke Location in Horizontal Plane and its Shape is Indicated Herein) by various Antenna Designs viz. (a) Antenna-A, (b) Antenna-B, (c) Antenna-C, (d) Antenna-D, and (e) Antenna-E.

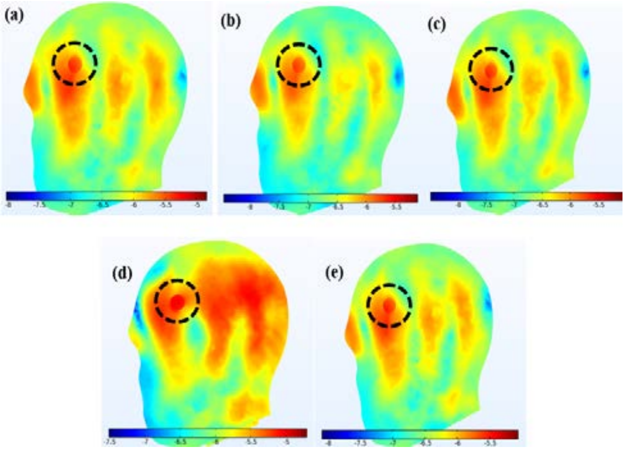


Fig. 6. Front view Layout of the Selected Local SAR Value Log-Scale Slice Plot showing the Exact Stroke Location (the Exact Positioning of the Stroke Location in Vertical Plane and its Shape is indicated herein) by various Antenna Designs viz. (a) Antenna-A, (b) Antenna-B, (c) Antenna-C, (d) Antenna-D, and (e) Antenna-E.

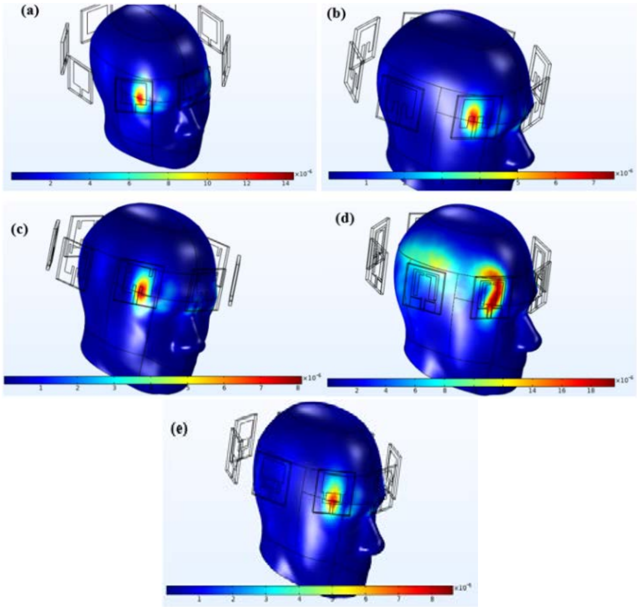


Fig. 7. Contour Plots representing Specific Absorption Rate in W/kg under the Influence of Microwave Irradiation by various Antenna Designs viz. (a) Antenna-A, (b) Antenna-B, (c) Antenna-C, (d) Antenna-D, and (e) Antenna-E.

The reduction in SAR values can be possibly due to two reasons viz. reduction of the input signal power or covering the antenna with the more dielectric material. By comparing the make designs of antenna-B, antenna-C, and antenna-E with antenna-A & antenna-D, the former one's had more dielectric material as the patch board. Hence, resulting in lower SAR values and making them a favorable choice of selection. Further, the antenna designs are also simulated to observe the effect on the reflection parameter/gain and the changes of SAR value with different designs. The investigated results are summarized for the reflection parameter and are separately discussed later in the last part of this section.

Fig. 8 shows the magnitude of the induced electric field within the head phantom with stroke included within, for all the antenna setups in terms of the x-component, y-component, and z-component. Herein the component of the electric field on the phantom surface is estimated and utilized for completing the imaging. The induced electric field values (emw) were 0.16, 0.1, 0.12, 0.25, and 0.13 for when the antenna-A, antenna-B, antenna-C, antenna-D, and antenna-E respectively are used for the stroke monitoring. The higher induced field implies that the details of the anatomical significance can be captured rapidly and with finer details, with measurement possible for minutely detailed examinations [42]. Hence, from this result, it comes out that the antenna-D should be the most desirable choice.

The performance comparison of the various antenna designs can be done by analyzing the 2-D far-field radiation patterns as shown in Fig. 9. This is a plot of the far-field norm represented in V/m. The 8x8 patch antenna array gain pattern (red) for all antennas show the main lobe and even the directional side lobes, but the main thing observed is that the emerging side lobes at 45° towards the right are more prominent and hence will dictate the direction of propagation. Though the 45° prominent lobes also signify that more area can be covered by the antenna designs. Considering the single patch antenna gain plot (blue), the pattern is similar for the antenna-B, antenna-C, and antenna-E; but for the antenna-A and antenna-D they have a distinctive bulging shape. The antenna-A shows very big area of the plot as compared to all others, representing that it can generate the wave with the most power, but it lacks directional streamlining. Hence, considering the directional antenna traits, other antenna designs will perform better.

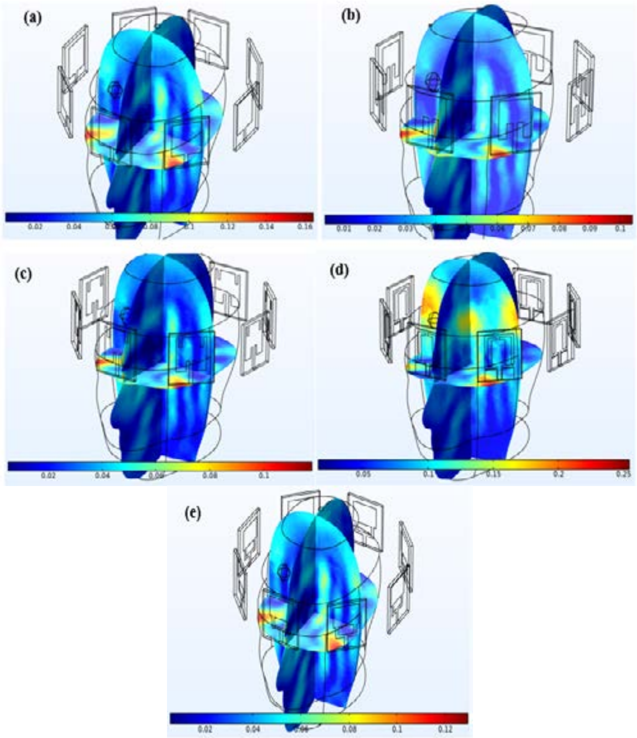


Fig. 8. The Induced Electric Field (emw) in the Stroke Containing Human Head Phantom under the Influence of Microwave Irradiation by various Antenna Designs viz. (a) Antenna-A, (b) Antenna-B, (c) Antenna-C, (d) Antenna-D, and (e) Antenna-E.

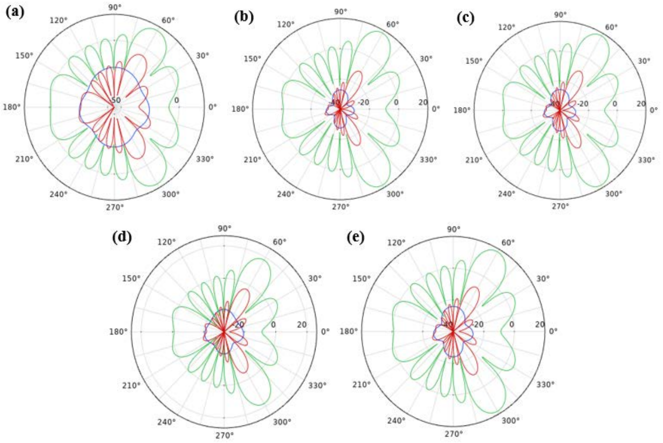


Fig. 9. 2-D Far-Field Radiation Patterns for the various Antenna Designs Studied in this Simulation Study viz. (a) Antenna-A, (b) Antenna-B, (c) Antenna-C, (d) Antenna-D, and (e) Antenna-E. Wherein, the Blue, Green and Red Trajectories represent Single Patch Antenna Gain, 8x8 uniform Array Factor, and 8x8 Patch Antenna Array Gain respectively.

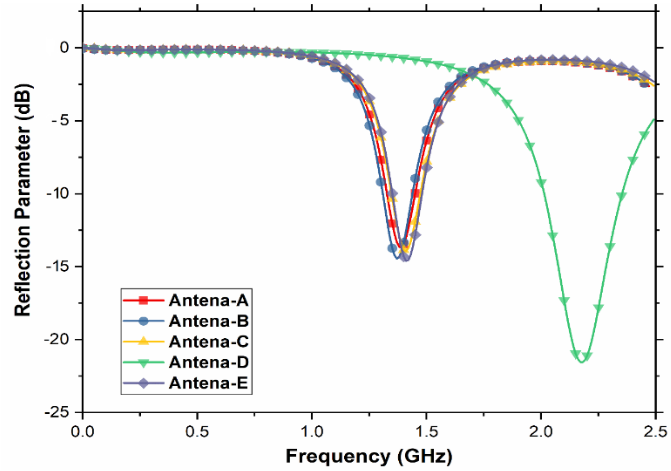


Fig. 10. Reflection Parameter (S11) Comparison Plot for all the Antenna Designs studied herein.

The reflection parameters also commonly known as reflection loss or simply S11 values; were calculated by frequency sweep in the range between 0 – 2.5 GHz for the five types of antennas and the comparison plot is shown in Fig. 10. All the values are lower than 0 dB (negative range) in the studied frequency range, which suggests that the antennas can operate efficiently when mounted on the head phantom [43]. The antennas A, B, C and E showed almost similar performance, with resonances around 1.3 - 1.5 GHz; the antenna D design, shows an S11 value under −22 dB for frequencies above 2.2 GHz. The antenna D operates more efficiently above 1.5 GHz, with a deep resonance around 2.2 GHz, making it ideal for the studied application in this research for stroke detection at the medically prescribed frequency level of 2.4 – 2.5 GHz.

IV. Conclusions

In this study, the concept of stroke identification and location is presented along with the analysis of the most suitable antenna design to be used, amongst the five chosen designs. The simulation study for the stroke identification with the five chosen antenna designs was successfully presented at 2.45 GHz frequency; as a means for addressing the queries related to the suitable antenna selection for the future experimental study. The results from the bioheat transfer module gave a clear understanding of the surface and sub-surface temperature rise due to the induced field, suggesting that the antenna-D had the highest sub-surface rise of 133.92 x 10^{-8} °C. But the main purpose of identifying the stroke presence with its shape, size, and location was accomplished by the SAR modeling for all the antenna design cases, by analyzing the appropriate horizontal and vertical SAR slices. Furthermore, the key concern of reducing the SAR for presenting the possibility of multiple usages without harmful effects unlike ionizing radiations was achieved with the antenna-A and antenna-D designs likewise. Overall, considering all the results discussed the antenna-D can be considered as the most suitable choice for future experimental prototyping and proving the efficacy of this concept. Based on this the experimental testing will be following as the next study.

Acknowledgment

The preferred work was supported by the Deanship of Scientific Research at Majmaah University under Project No. RGP- 2019-32.

References

[1] World Health Organization Global health estimates: deaths by cause, age, sex and country, 2000-2012. Geneva, WHO 2014, 9.

[2] Owolabi, M.O.; Akarolo-Anthony, S.; Akinyemi, R.; Arnett, D.; Gebregziabher, M.; Jenkins, C.; Tiwari, H.; Arulogun, O.; Akpalu, A.; Sarfo, F.S.; Obiako, R.; Owolabi, L.; Sagoe, K.; Melikam, S.; Adeoye, A.M.; Lackland, D.; Ovbiagele, B.; Members of the H3Africa Consortium The burden of stroke in Africa: a glance at the present and a glimpse into the future. Cardiovasc J Afr 2015, 26, S27-38.

[3] Johnson, W.; Onuma, O.; Owolabi, M.; Sachdev, S. Stroke: a global response is needed. Bull World Health Organ 2016, 94, 634-634A.

[4] Feigin, V.L.; Forouzanfar, M.H.; Krishnamurthi, R.; Mensah, G.A.; Connor, M.; Bennett, D.A.; Moran, A.E.; Sacco, R.L.; Anderson, L.; Truelsen, T.; O'Donnell, M.; Venketasubramanian, N.; Barker-Collo, S.; Lawes, C.M.; Wang, W.; Shinohara, Y.; Witt, E.; Ezzati, M.; Naghavi, M.; Murray, C.; Global Burden of Diseases, Injuries, and Risk Factors Study 2010 (GBD 2010) and the GBD Stroke Experts Group Global and regional burden of stroke during 1990-2010: findings from the Global Burden of Disease Study 2010. Lancet 2014, 383, 245-254.

[5] Strong, K.; Mathers, C.; Bonita, R. Preventing stroke: saving lives around the world. The Lancet Neurology 2007, 6, 182-187.

[6] Chandra, R.; Zhou, H.; Balasingham, I.; Narayanan, R.M. On the opportunities and challenges in microwave medical sensing and imaging. IEEE transactions on biomedical engineering 2015, 62, 1667-1682.

[7] Walsh, K.B. Non-invasive sensor Technology for Prehospital Stroke Diagnosis: current status and future directions. International Journal of Stroke 2019, 14, 592-602.

[8] Shao, Y.; Tsai, K.; Kim, S.; Wu, Y.; Demissie, K. Exposure to Tomographic Scans and Cancer Risks. JNCI Cancer Spectrum 2020, 4, pkz072.

[9] Mayosi, B.M.; Lawn, J.E.; Van Niekerk, A.; Bradshaw, D.; Karim, S.S.A.; Coovadia, H.M.; Lancet South Africa team Health in South Africa: changes and challenges since 2009. The Lancet 2012, 380, 2029-2043.

[10] O'donnell, M.J.; Xavier, D.; Liu, L.; Zhang, H.; Chin, S.L.; Rao-Melacini, P.; Rangarajan, S.; Islam, S.; Pais, P.; McQueen, M.J. Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22

countries (the INTERSTROKE study): a case-control study. The Lancet 2010, 376, 112-123.

[11] A. Zamani, A. M. Abbosh, and A. T. Mobashsher, “Fast frequency-based multistatic microwave imaging algorithm with application to brain injury detection,” IEEE Trans. Microw. Theory Techn., vol. 64, no. 2, pp. 653– 662, Feb. 2016.

[12] Walsh, K.B. Non-invasive sensor Technology for Prehospital Stroke Diagnosis: current status and future directions. International Journal of Stroke 2019, 14, 592-602.

[13] Mobashsher, A.T.; Bialkowski, K.; Abbosh, A.; Crozier, S. Design and experimental evaluation of a non-invasive microwave head imaging system for intracranial haemorrhage detection. Plos one 2016, 11, e0152351.

[14] Merunka, I.; Massa, A.; Vrba, D.; Fiser, O.; Salucci, M.; Vrba, J. Microwave tomography system for methodical testing of human brain stroke detection approaches. International Journal of Antennas and Propagation 2019.

[15] Fhager, A.; Candefjord, S.; Elam, M.; Persson, M. Microwave diagnostics ahead: Saving time and the lives of trauma and stroke patients. IEEE Microwave Magazine 2018, 19, 78-90.

[16] Karadima, O.; Rahman, M.; Sotiriou, I.; Ghavami, N.; Lu, P.; Ahsan, S.; Kosmas, P. Experimental Validation of Microwave Tomography with the DBIM-TwIST Algorithm for Brain Stroke Detection and Classification. Sensors 2020, 20, 840.

[17] Mobashsher, A.T.; Abbosh, A. On-site rapid diagnosis of intracranial hematoma using portable multi-slice microwave imaging system. Scientific reports 2016, 6, 37620.

[18] Hopfer, M.; Planas, R.; Hamidipour, A.; Henriksson, T.; Semenov, S. Electromagnetic Tomography for Detection, Differentiation, and Monitoring of Brain Stroke: A Virtual Data and Human Head Phantom Study. IEEE Antennas and Propagation Magazine 2017, 59, 86-97.

[19] Bisio, I.; Fedeli, A.; Lavagetto, F.; Pastorino, M.; Randazzo, A.; Sciarrone, A.; Tavanti, E. A numerical study concerning brain stroke detection by microwave imaging systems. Multimedia Tools Appl 2018, 77, 9341-9363.

[20] Alqadami, A.S.; Bialkowski, K.S.; Mobashsher, A.T.; Abbosh, A.M. Wearable electromagnetic head imaging system using flexible wideband antenna array based on polymer technology for brain stroke diagnosis. IEEE transactions on biomedical circuits and systems 2018, 13, 124-134.

[21] Afsari, A.; Abbosh, A.M.; Rahmat-Samii, Y. Modified born iterative method in medical electromagnetic tomography using magnetic field fluctuation contrast source operator. IEEE Trans Microwave Theory Tech 2018, 67, 454-463.

[22] Maffongelli, M.; Poretti, S.; Salvadè, A.; Monleone, R.; Pagnamenta, C.; Fedeli, A.; Pastorino, M.; Randazzo, A. Design and experimental test of a microwave system for quantitative biomedical imaging, 2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA), IEEE: 2018; , pp. 1-6.

[23] Persson, M.; Fhager, A.; Trefná, H.D.; Yu, Y.; McKelvey, T.; Pegenius, G.; Karlsson, J.; Elam, M. Microwave-based stroke diagnosis making global prehospital thrombolytic treatment possible. IEEE Transactions on Biomedical Engineering 2014, 61, 2806-2817.

[24] Candefjord, S.; Winges, J.; Malik, A.A.; Yu, Y.; Rylander, T.; McKelvey, T.; Fhager, A.; Elam, M.; Persson, M. Microwave technology for detecting traumatic intracranial bleedings: tests on phantom of subdural hematoma and numerical simulations. Med Biol Eng Comput 2017, 55, 1177-1188.

[25] D. Ireland, K. Bialkowski, and A. Abbosh, “Microwave imaging for brain stroke detection using Born iterative method,” IET Microw. Antennas Propag., vol. 7, no. 11, pp. 909–915, Aug. 2013.

[26] IEEE Standards Coordinating Committee 34 IEEE Recommended Practice for Determining the Peak Spatial-average Specific Absorption Rate (SAR) in the Human Head from Wireless Communications Devices: Measurement Techniques, Institute of Electrical and Electronic Engineers: 2003.

[27] IEC 62209–1 Human exposure to radio frequency fields from hand-held and body-mounted wireless communication devices—Human models, instrumentation, and procedures—Part 1: Procedure to determine the specific absorption rate (SAR) for hand-held devices used in close proximity to the ear (frequency range of 300 MHz to 3 GHz). 2005.

[28] ARIB, S. Specific absorption rate (SAR) estimation for cellular phone. ARIB STD-T56 1998.

[29] ENS, C.S. 50361:“Basic Standard for the measurement of Specific Absorption Rate related to human exposure to electromagnetic fields from mobile phones (300 MHz-3 GHz),”. Brussels, Belgium, CENELEC 2001.

[30] Fields, R.E. Evaluating compliance with FCC guidelines for human exposure to radiofrequency electromagnetic fields. OET bulletin 1997,

[31] Levoy, M. The Stanford volume data archive. URL: http://graphics.stanford.edu/data/voldata 2001.

[32] Schmid, G.; Neubauer, G.; Mazal, P.R. Dielectric properties of human brain tissue measured less than 10 h postmortem at frequencies from 800 to 2450 MHz. Bioelectromagnetics: Journal of the Bioelectromagnetics Society, The Society for Physical Regulation in Biology and Medicine, The European Bioelectromagnetics Association 2003, 24, 423-430.

[33] Yang, Z.; Xiao, S.; Zhu, L.; Wang, B.; Tu, H. A circularly polarized implantable antenna for 2.4-GHz ISM band biomedical applications. IEEE Antennas and Wireless Propagation Letters 2017, 16, 2554-2557.

[34] Chen, M.M.; Holmes, K.R. Microvascular contributions in tissue heat transfer. Ann N Y Acad Sci 1980, 335, 137-150.

[35] Weinbaum, S.; Jiji, L. A new simplified bioheat equation for the effect of blood flow on local average tissue temperature. 1985.

[36] Baish, J.; Ayyaswamy, P.; Foster, K. Small-scale temperature fluctuations in perfused tissue during local hyperthermia. 1986.

[37] Perez, F.P.; Bandeira, J.P.; Morisaki, J.J.; Krishna Peddinti, S.V.; Salama, P.; Rizkalla, J.; Rizkalla, M.E. Antenna Design and SAR Analysis on Human Head Phantom Simulation for Future Clinical Applications. J Biomed Sci Eng 2017, 10, 421-430.

[38] Shih, T.; Yuan, P.; Lin, W.; Kou, H. Analytical analysis of the Pennes bioheat transfer equation with sinusoidal heat flux condition on skin surface. Med Eng Phys 2007, 29, 946-953.

[39] Bond, S.; Wang, K. The Impact of Cell Phone Towers on House Prices in Residential Neighborhoods. Appraisal J 2005, 73.

[40] El-Sharkawy, A.M.; Qian, D.; Bottomley, P.A.; Edelstein, W.A. A multichannel, real-time MRI RF power monitor for independent SAR determination. Med Phys 2012, 39, 2334-2341.

[41] Lin, J.C. A new IEEE standard for safety levels with respect to human exposure to radio-frequency radiation. IEEE Antennas and Propagation Magazine 2006, 48, 157-159.

[42] Qureshi, A.M.; Mustansar, Z. Levels of detail analysis of microwave scattering from human head models for brain stroke detection. PeerJ 2017, 5, e4061.

[43] Guo, W.; Ahsan, S.; He, M.; Koutsoupidou, M.; Kosmas, P. Printed monopole antenna designs for a microwave head scanner, 2018 18th Mediterranean Microwave Symposium (MMS), IEEE: 2018; , pp. 384-386

Verification of Himawari-8 Observation Data using Cloud Optical Thickness (COT) and Cloud Image Energy

Umar Ali Ahmad[1], Alex Lukmanto Suherman[5]
Agus Virgono[8], Burhanuddin Dirgantoro[9]
Reza Rendian Septiawan[10]
Department of Computer Engineering, School of Electrical Engineering, Telkom University, Bandung, Indonesia

Wendi Harjupa[2]
Center of Atmospheric Science and Technology
LAPAN, Bandung, Indonesia

Dody Qory Utama[3]
School of Informatics. Telkom University
Bandung, Indonesia

Risyanto[4], Prayitno Abadi[7]
Space Science Center. LAPAN
Bandung, Indonesia

Wahyu Pamungkas[6]
Fakultas Teknik Telekomunikasi dan Elektro
Institut Teknologi Telkom
Purwokerto, Indonesia

Mas'ud Adhi Saputra[11]
Center for Material and Technical Products
Ministry of Industry, Jakarta
Indonesia

***Abstract*—Himawari-8 satellite cloud observation data covers all areas of Indonesia. The cloud observation data can be used for observations of current weather conditions and short-term predictions. This paper reports the verification method of Himawari-8 Observation Data using Cloud Optical Thickness (COT) and compared to Cloud Image Energy. The verification test was carried out to determine the accuracy of Himawari-8's observations. COT data were verified using energy data from the observation image of the time-lapse camera. First, the time-lapse camera captures and classifies the cloud image. Subsequently, the energy of each image frame was calculated and re-grouped the result based on the energy to determine the type of the cloud. The results show that there is a positive correlation between COT and low energy values with cumulonimbus cloud detection, on the contrary for Cirrus-cloud type. However, the data requires a more accurate observation method to obtain data from cloud images on the Himawari-8 satellite, specifically for regions with a small spatial size of 4 km and thin clouds in the lower layer.**

Keywords—Himawari-8; COT; image classification; cloud energy

I. Introduction

Rain formation is not an easy phenomenon to understand, although it can be simplified. Rain can occur from the process of evaporation, cloud formation (condensation), and the dropping of water from the atmosphere (precipitation). However, in fact, the process of rain formation becomes complicated as well as the process of forming clouds. A combination of atmospheric parameters, such as temperature, pressure, and humidity at a certain value, can produce clouds.

The next question is whether the formed clouds will certainly produce rain or not. Weather prediction, such as the prediction of extreme rain in Indonesia, is a very challenging matter. Indonesia's geographic location, which is located between two continents (Asia and Australia) and two oceans (Indian and Pacific), causes Indonesia to be traversed by many global circulations. For example, the Madden-Julian Oscillation (MJO) and Monsoon circulations. In addition, Indonesia's hilly topography causes local circulation, such as convection to be very active, which also has an impact on the weather.

Recent progress in the satellite cloud technologies, such as Himawari-8, enable us to predict the current weather and short-term conditions. Himawari-8 is a geostationary weather satellite that launched by the Japan Meteorology Agency (JMA) in 2014. Himawari-8 is equipped with the Advanced Himawari Imager (AHI) instrument with a significant increase in the number of radiometric spectral channels and spatial resolution, compared to its previous generations [1]. The number of AHI observation channels consists of three Visible channels, three Near Infra-Red (NIR) channels, and 10 Infra-Red (IR) channels, with a spatial resolution of 0.5 and 1 km for Visible and NIR, and 2 km for IR in Japan [1].

Research using AHI has also been applied by Letu *et al*. [2] by using the Look-Up Table (LUT) method to calculate high temporal (10 minutes) and high spatial (5 km) of surface solar radiation from AHI cloud properties. AHI properties, especially those related to ice cloud products, were also used by Letu *et al*. [2] to monitor the growth process, including monitoring variations of cloud properties and deposition in Deep Convective (DC) clouds. Research conducted by Lee *et al*. [3] also uses AHI, which is sensitive to clouds and gas absorption, to improve the algorithm for calculating Outgoing Longwave Radiation (OLR) at the peak of the atmosphere.

In Indonesia, the Himawari-8 satellite observations cover the entire area from Western to Eastern parts of Indonesia with time and space resolutions of 10 minutes and 4 km. Several researches using Himawari-8 satellite [4][5] are used to detect the condition of volcanoes in Indonesia. Another application of the Himawari-8 satellite in Indonesia was carried out by Osawa *et al*. [6], researching Long Internal Solitary Waves (ISWs), which seem to be generated from the Sape Strait, which then pro-gates through the Java Sea, the Sumba Strait, and the Savu Sea. In a country directly adjacent to Indonesia, namely Papua, Himawari-8 is also used to investigate the topography of the country's territory [7]. These reported researches indicate the advantages of using the Himawari-8 satellite to monitor cloud activity near real-time.

The data verification method is crucial to maximize the use of the Himawari-8 Satellite data. One example of the use of Himawari-8 data is for rain estimation. Cloud Optical Thickness (COT) can provide qualitative cloud thickness information. Concerning the decrease in solar radiation, COT has an exponential value, where COT and solar radiation data were taken from Nanyang Technological University (NTU) for four months [8].

Research that examines the effect of COT, the albedo of ground surface, and the above-cloud absorbing dust layer structure result in the phenomenon that higher reflectance is polarized along with the increase in COT and will be saturated when COT = 10 [9]. Meanwhile, the Cloud Optical Thickness-Retrieval Over Snow (CROS) algorithm has been used in snowy areas [10] to tackle this issue, but the issue of estimation accuracy is crucial. Information on cloud thickness is important in estimating rainfall [11]. Sakai *et al*. [12] used cloud observations from the Himawari-8 satellite to produce COT. Harjupa *et al*. [13] used cloud data from the Himawari-8 to predict the extreme rain within the next hour (nowcasting). However, Himawari-8 satellite cannot observe multi-layered clouds and thus the accuracy of COT and predictive products becomes our concern. The aim of this paper is to tackle this limitation. For this reason, COT verification is carried out with data from camera observations from the earth's surface. In this paper, we are using a lapse-time camera which can take image of clouds with high time and spatial resolution. Observations from the earth's surface are needed to observe clouds in the lower layers.

This paper consists of the following sections. In Section II, we define our method to observe the cloud properties and illustrate the process of classification and detection using the time-lapse camera. In Section III, we present the cloud grouping from a time-lapse camera based on the energy level, compared with the images from the satellite. In Section IV, we discussed the results presented in the previous chapter in a structured manner. In Section V, the results are concluded and further research plans are presented.

II. Materials and Methods

Fig. 1 shows cloud detection method using the Himawari-8 satellite, which is carried out from space and cloud detection using a camera from the earth's surface. The results suggest that the observations from the satellites can only detect clouds in the upper layers, not the lower layers. Camera observations made from the earth's surface will complement the cloud observation data from the Himawari-8 satellite.

The results from cloud observation using a Himawari-8 satellite will be derived in the form of a COT product. The cloud observation image will be analyzed and the energy value of the cloud will be calculated. The two observations was performed and compared to determine the accuracy of the Himawari-8 observations. Fig. 2 shows data verification diagram for the Himawari-8 and the camera.

A. Cloud Top Layer Thickness

The COT value in this study was obtained from the difference in the value of brightness temperature (BT) of band13 and brightness temperature difference (BTD) of band15 observed by Himawari-8. The height of the cloud can be measured qualitatively using band13, where the lower the brightness of the temperature band13 indicates the high of the cloud peak due to the reduced surface radiation value. Nishi *et al* [14] found that at BT value of 220 K, the cloud height is about 14.4 km. By comparing the BT value between band13 and band15, the COT value can be determined. The COT value will be bigger at the higher cloud positions [12].

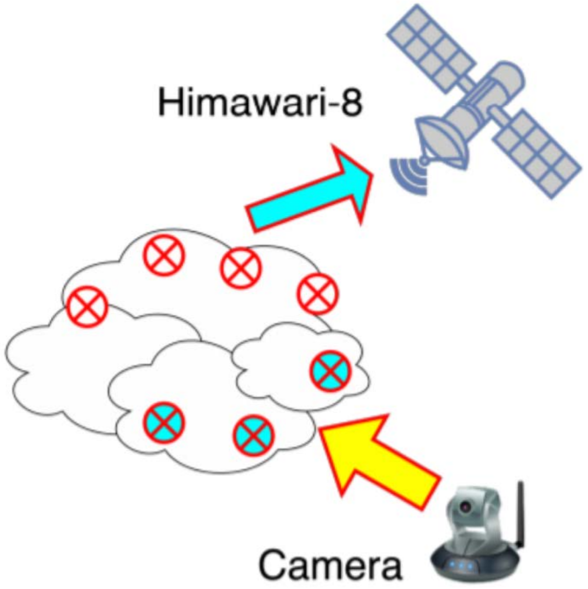


Fig. 1. Cloud Observation using the Himawari-8 Satellite and a Time-Lapse Camera.

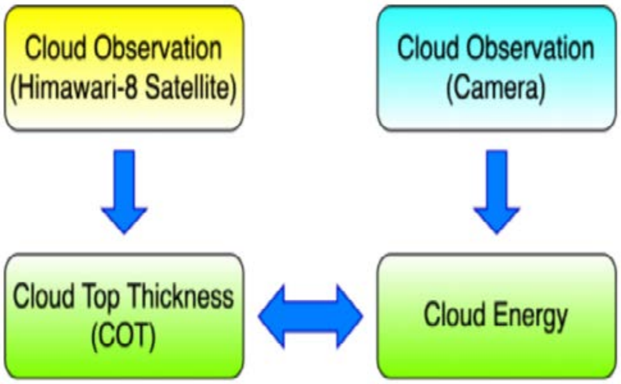

Cloud Observation (Himawari-8 Satellite)
Cloud Observation (Camera)
Cloud Top Thickness (COT)
Cloud Energy

Fig. 2. Comparison of COT and Cloud Energy.

The COT value, which is the difference in the BT value of the two bands (BTD), can be used to classify the types of rain clouds [15]. The cloud types of the two BT are shallow and non-shallow clouds. With these results, the COT value will be beneficial for estimating the rain generated by clouds. In this study, the location and time of cloud detection were selected manually, namely by matching the same time and place with the cloud observed from the time-lapse camera.

B. Classification and Cloud Detection using Time-Lapse Camera

Fig. 3 shows how we classify the clouds from time-lapse camera. First step is converting the time-lapse video images into several images, which will be analyzed and classified. Following this, the energy of each frame of the image were calculated. Subsequently, each cloud was classified based on the energy value obtained for each frame.

The framing process in Fig. 4 aims to take a sample of time unity from the existing time-lapse video. This process was carried out since the movement of the cloud is slow enough for the unity of time. It can be concluded that the near union time can be ascertained that the cloud will have adjacent energy. In this study, the framing process was performed by taking one digital image frame every ten frames, with 30 frames per second.

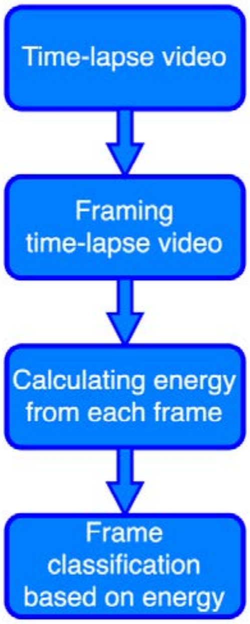

Time-lapse video
Framing time-lapse video
Calculating energy from each frame
Frame classification based on energy

Fig. 3. The Cloud Classification Method using a Time-Lapse Camera.

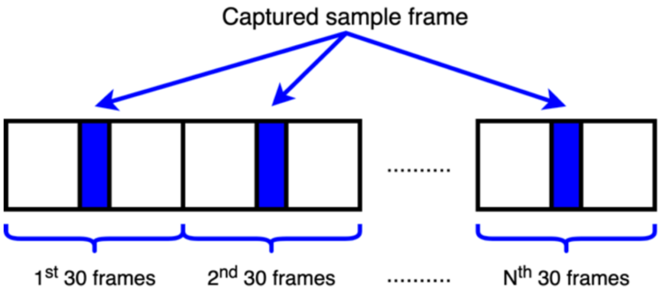


Fig. 4. Video Framing Method Lapse-Time.

The average brightness of each frame obtained from the framing process was used to calculate the energy levels. This process can be seen in Fig. 5.

The frame (a digital image) will indicates the brightness value for each pixel, then add up all the brightness values and divided by the calculated pixel count, as shown in the following mathematical notation:

$$\textbf{Frame energy} = \frac{\sum_{x=0}^{n} pixel(x)}{n}$$

After getting the energy level for each frame, the frames will be grouped according to seven energy values, namely group 1 (energy of 0-20), group 2 (energy of 20-40), group 3 (energy of 40-60), group 3 (energy of 60-80), group 4 (energy of 80-100), group 5 (energy of 100-120), group 6 (energy of 120-140), and group 7 (energy above 140). The energy level will determine the type of cloud. Fig. 6 shows the grouping of frames based on energy levels.

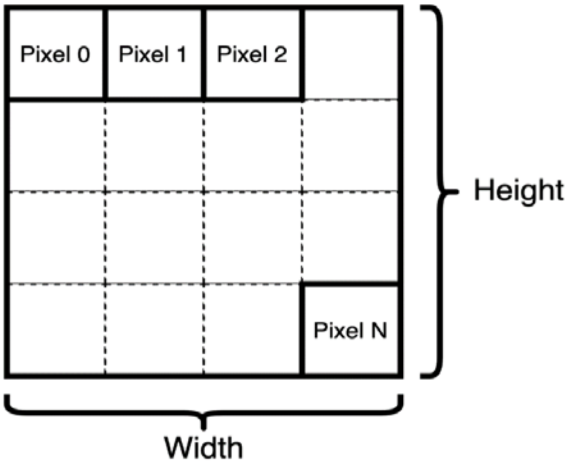


Fig. 5. Frame Structure when Measuring its Energy Level.

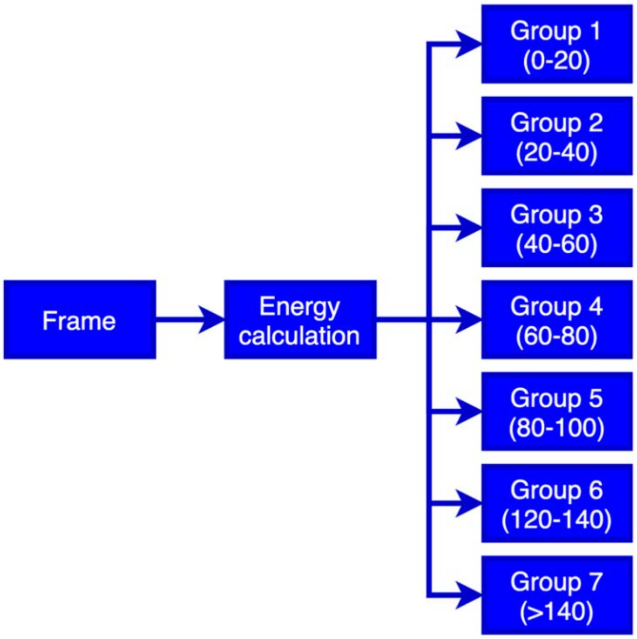
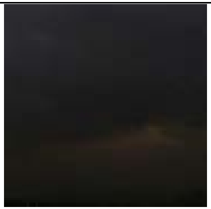
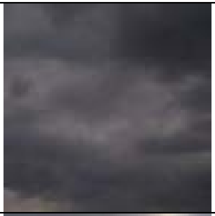
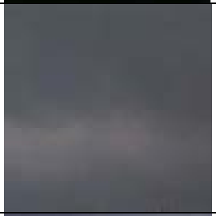
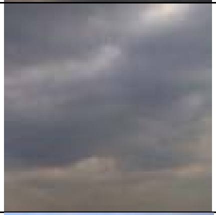
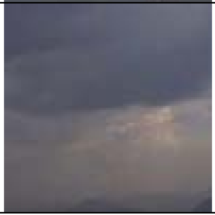
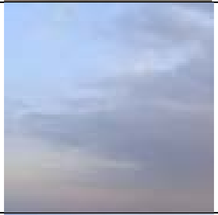
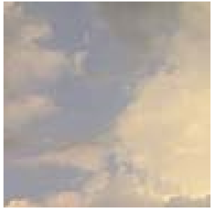
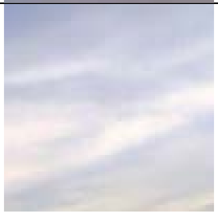
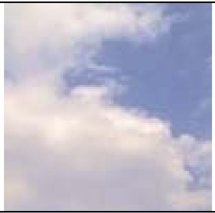
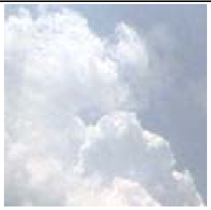


Fig. 6. Grouping of the Frame's Energy.

III. Results

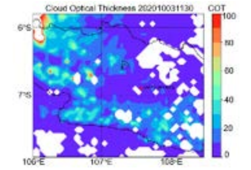
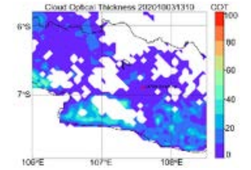
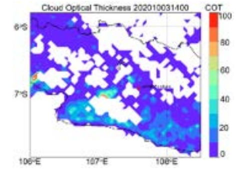
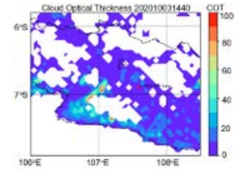
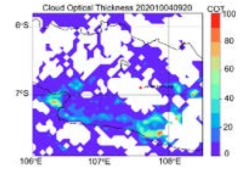
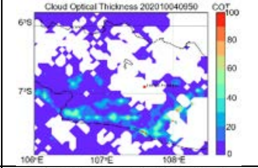
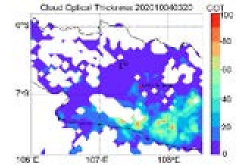
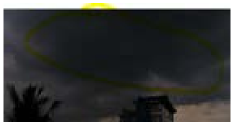
Table I shows the image resulted from the time-lapse camera, showing various sample images belong to the respective cloud energy groups. Comparison of COT values and cloud energy is shown.

TABLE I. The Results of Cloud Grouping from a Time-Lapse Camera based on the Energy Level

No.	Picture samples	
1		
2		
3		
4		
5		
6		
7		

The comparison of the COT value and cloud energy is shown in Table II. The COT value is taken in the yellow circle area, which is the same area as the cloud observed by the time-lapse camera.

TABLE II. The Results of Cloud Grouping from a Time-Lapse Camera Compared with the Picture from the Satellite

COT	Picture samples from satellite	Energy level	Picture samples from the camera
1.16	 Cloud Optical Thickness 202010031130 COT 6°S 7°S 106°E 107°E 108°E 100 80 60 40 20 0	138	
3.35	 Cloud Optical Thickness 202010031310 COT 6°S 7°S 106°E 107°E 108°E 100 80 60 40 20 0	110	
0.55	 Cloud Optical Thickness 202010031400 COT 6°S 7°S 106°E 107°E 108°E 100 80 60 40 20 0	126	
9.31	 Cloud Optical Thickness 202010031440 COT 6°S 7°S 106°E 107°E 108°E 100 80 60 40 20 0	161	
nan	 Cloud Optical Thickness 202010040920 COT 6°S 7°S 106°E 107°E 108°E 100 80 60 40 20 0	153	
nan	 Cloud Optical Thickness 202010040950 COT 6°S 7°S 106°E 107°E 108°E 100 80 60 40 20 0	213	
1.99	 Cloud Optical Thickness 202010040320 COT 6°S 7°S 106°E 107°E 108°E 100 80 60 40 20 0	51	

IV. Discussions

A. Energy-based Grouping of Clouds from Time-Lapse Camera

In group 1, with an energy value of 0-20, the cloud conditions are very dark. It was found that the value of cloud energy becomes very small at sunset or night, as shown in Table I. From the grouping clouds data based on this energy, it is found that clouds with different types can produce the same energy as seen in the cloud, with energy group 5, 6 and 7. Types of clouds in the energy group 5 are Stratus and Cirrus clouds. The same type of clouds is found in group 6 and 7. In addition, the same type of clouds also found from the Himawari-8 observations, where a small value of brightness temperature (BT) could be produced by Cirrus, Stratus or Cumulus-type clouds.

B. Cloud Energy from the Camera and COT of Himawari-8

COT values and cloud energy in group 1 in Table II is 1.16 and 138, respectively. Since the found energy level is 138, it is a Stratus-type cloud and not too thick. Clouds with lower energy values are seen in the comparison of the COT value and cloud energy level of group 2 in Table II. The initial thickness makes the detected energy lower, namely, 110, and this value is comparable to the COT value, which is higher than the COT value in comparison with number 1. Low energy and COT values indicate a thick cloud that may produce a heavy rain. In comparison, number 3 in the same figure, it can be seen that the COT value is very low, namely 0.55, with a high energy value of 126. The type of cloud in the ratio number 3 are tall and thin clouds, like Cirrus-cloud type. In comparison, number 4 shows a difference where the high COT value is 9.31, but the energy value is also high, namely 161. The difference in COT values results an error during the prediction of heavy rain. With the help of time-lapse camera, the results of cloud observations from the Himawari-8 satellite can be validated. In comparison between number 5 and 6, no COT value is detected, while the camera has energy values, namely 153 and 213. From the image of the cloud that is extracted, the energy shows that it is spatially and the observed cloud area is not very large. This is a weakness of Himawari-8's observations, as it cannot observe clouds with a spatial area smaller than 4 km. This result shows the important aspect of using a time-lapse camera, as it can detect clouds with a smaller spatial resolution. In comparison, number 7 shows that the COT value is not too high with a low energy value, suggesting the cloud type is a thin cloud or Cirrus cloud-type.

V. Conclusions

In this initial verification of Himawari-8 observation data and cloud energy from camera observations, several conclusions were obtained: (a) high COT values and low energy values indicate cumulonimbus cloud detection, on the contrary for cirrus cloud-type, the energy value will be high with low COT, (b) the Himawari-8 satellite cannot capture cloud images with a small spatial size of 4 km and thin clouds in the lower layer, (c) it is necessary to establish a relationship between the COT value and the energy value of each cloud type to tackle the limitation of Himawari-8 observation data and to increase the capability or accuracy of derivative product such rain prediction using Himawari-8's observation data.

VI. Future Works

Further experimental studies are required for validation of the proposed method. The increasing dataset will have the advantage to increase the result validity throughout a longer period of observation. Furthermore, the relation information must be expanded to other research fields, such as weather prediction.

Acknowledgments

This research was supported by Telkom University, the Center of Atmospheric Science and Technology of LAPAN, and the Space Science Center of LAPAN, Indonesia.

References

[1] K. Bessho, et al., “An introduction to Himawari-8/9 — Japan’s new-generation geostationary meteorological satellites,” J. Meteorol. Soc. Japan. Ser. II, 94(2), 2016, pp. 151–183. DOI: 10.2151/jmsj.2016-009.

[2] H. Letu, et al., “Ice Cloud Properties From Himawari-8/AHI Next-Generation Geostationary Satellite: Capability of the AHI to Monitor the DC Cloud Generation Process,” IEEE Trans. Geosci. Remote Sens., 57(6), 2019, pp. 3229–3239. DOI: 10.1109/TGRS.2018.2882803.

[3] B. Y. Kim and K. T. Lee, “Using the Himawari-8 AHI multi-channel to improve the calculation accuracy of outgoing longwave radiation at the top of the atmosphere,” Remote Sens., 11(5), 2019. DOI: 10.3390/rs11050589.

[4] F. Marchese, et al., “Monitoring the Agung (Indonesia) ash plume of November 2017 by means of infrared Himawari 8 data,” Remote Sens., 10(6), 2018. DOI: 10.3390/rs10060919.

[5] T. Kaneko, et al., “Himawari-8 infrared observations of the June–August 2015 Mt Raung eruption, Indonesia,” Earth Planets Space. Springer Berlin Heidelberg, 70(1), 2018, pp. 1–9. DOI: 10.1186/s40623-018-0858-9.

[6] I. W. G. A. Karang, Chonnaniyah, and T. Osawa, “Internal solitary wave observations in the Flores Sea using the Himawari-8 geostationary satellite,” Int. J. Remote Sens.. Taylor & Francis, 41(15), 2020, pp. 5726–5742. DOI: 10.1080/01431161.2019.1693079.

[7] H. Iwabuchi et al., “Cloud property retrieval from multiband infrared measurements by Himawari-8,” J. Meteorol. Soc. Jpn., 96B(September 2017), 2018, pp. 27–42. DOI: 10.2151/jmsj.2018-001.

[8] F. Yuan et al., “Correlation between cloud optical thickness and solar radiation,” 2016 USNC-URSI Radio Science Meeting (Joint with AP-S Symposium), USNC-URSI 2016 - Proceedings, (1), 2016, pp. 105–106. DOI: 10.1109/USNC-URSI.2016.7588534.

[9] H. Shang, et al., "The effect of cloud optical thickness, ground surface albedo and above-cloud absorbing dust layer on the cloudbow structure," 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, 2016, pp. 2174-2176, DOI: 10.1109/IGARSS.2016.7729561.

[10] C. Schlundt, et al., “Determination of cloud optical thickness over snow using satellite measurements in the oxygen A-band,” IEEE Geosci. Remote. Sens. Lett., 10(5), 2013, pp. 1162–1166. DOI: 10.1109/LGRS.2012.2234720.

[11] S. Chakraborty and A. Maitra, “Interrelation between microphysical and optical properties of cloud and rainfall in the Indian region,” Indian J. Radio and Space Phys., 42, 2013, pp. 105-112.

[12] Sakai, et al., “Development of a Rapid Retrieval Method for Cloud Optical Thickness and Cloud-Top Height using Himawari-8 Infrared Measurements,” Sci. Online Lett. Atmosphere, 15, 2019, pp. 57–61.

[13] W. Harjupa, et al., “Fundamental Investigation of Generation of Guerilla-heavy rainfall using Himawari-8 and XRAIN information in Kinki region,” J. Hydraul. Eng., JSCE, 74(4), 2017, pp. 283–288.

[14] A. Hamada and N. Nishi, "Development of a Cloud-Top Height Estimation Method by Geostationary Satellite Split-Window Measurements Trained with CloudSat Data." J. Appl. Meteor. Climatol., 49 (9), 2010, pp. 2035–2049.

[15] D. So and D. B. Shin, " Classification of precipitating clouds using satellite infrared observations and its implications for rainfall estimation," Advances in Remote Sensing of Rainfall and Snowfall, 2018. https://doi.org/10.1002/qj.3288.

A Smart Approximation Algorithm for Minimum Vertex Cover Problem based on Min-to-Min (MtM) Strategy

Jawad Haider[1], Muhammad Fayaz[2]
Department of Computer Science
University of Central Asia
Naryn 722918, Kyrgyzstan

Abstract—**In this paper, we have proposed an algorithm based on min-to-min approach. In the proposed algorithm first the degree of each vertex of the graph is calculated. Next the vertex with minimum degree is selected, after which all the neighbors of the minimum degree are located. In the neighbors of the minimum degree vertex, again the vertex with the minimum degree is found and put into the set minimum vertex cover and deleted from the graph. Again, the degree of each vertex of the updated graph is calculated and again the same process is repeated until the graph becomes empty. In case of tie, all the neighbors of the minimum degree vertices are computed and then the minimum degree vertex in all of them is added to minimum vertex degree set. The same process is repeated until the graph becomes empty. The proposed algorithm is a very simple, efficient, and easy to understand and implement. The proposed min-to-min algorithm is evaluated on small as well as on large benchmark instances and the results indicate that the performance of the min-to-min algorithm is far better as compared to the other state-of-art algorithms in term of accuracy and computation complexity. We have also used the proposed method to solve the maximum independent set problem.**

Keywords—Minimum vertex cover; approximation algorithms; maximum independent set; benchmark instances; graph theory

I. Introduction

A graph in field of computer science is the set of vertices, and collection of edges each of which connects a pair of vertices [1]. Edges are the links between the vertices. Mathematically, a graph is represented as G (V, E), where V denotes the set of vertices and E the collection of edges[2] .In graph theory a vertex cover is defined as the subset Vc $\subseteq$ V, such that the vertices in the subset Vc covers all the edges E in the graph G(V, E). Minimum vertex cover(MVC) is the minimum number of elements in the subset Vc which covers all the edges in the graph [3]. The minimum vertex cover problem is well known in graph theory because of its real-life applications in diverse fields. For example, MVC has application in civil and electrical engineering, map labeling, sensor networks as well as in very-large scale integration (VLSI) design, bioinformatics and biochemistry, protein sequencing and gene regulatory network [4-7]. A real-life example to better explain the application of MVC is the positions of guards in a museum. Where each edge of the graph represents the corridors of the museum and each vertex denotes the position of the guards. To station minimum guards while still covering all the corridors of the museum is minimum vertex cover problem [8].

Minimum vertex cover problem comes in two versions-optimization version and decision version. The decision version is a Boolean type problem. Where the question is to find, if there exists a solution of desired size k? k is the minimum number of vertices that should be used. And the result is either true or false (Yes or No). The optimization version is all about finding the optimal solution[5]. Our concern in this paper will be with second version i.e optimization version of MVC.

In 1979 S.Cook put two conditions for NP problems to be called as NP-complete. 1) It should be NP-hard and 2) it should be reducible to any NP-complete problem in deterministic polynomial time [9]. Cook put the Boolean Satisfiability (SAT) problem as the basic problem in the set of NP-complete problems. As clique can be converted into SAT problem in polynomial time, so clique belongs to SAT problems. Moreover, maximum independent set (MIS) problem can be converted to clique problem as well, similarly minimum vertex cover can be converted into MIS problem in polynomial time, hence the above problems are interchangeable and are considered as NP-complete problems [10]. To conclude, minimum vertex cover (MVC) problem is NP-complete problem. MIS, MC and MVC are all complimentary concepts and NP-complete problems.

Maximum independent set is also a great problem, where the optimizing task is to find a set containing maximum number of such vertices of the graph where no pair of the vertices share an edge or are adjacent to each other [11]. Its application can be found in map labelling problems where the names of adjacent cities should be placed nearest to the city without any overlap [7]. MIS is also used in high level synthesis and physical design automation, while computing the MIS in a graph can be used to determine the maximum number of processors required for parallel execution and also used in channel routing problems of physical design automation, for instance k-layer routing for Printed Circuit Boards and Multi-Chip Modules [12]. Moreover, MIS have application in information retrieval, classification theory, scheduling, economics, computer vision and experimental design [13].

There are two types of algorithms to solve any NP-complete problems. These are the approximation algorithms and exact algorithms [14,15]. The exact algorithms will always provide a solution that is optimal, but the computation time increases exponentially with the size of the problem. Therefore, the exact algorithms are best option for small size problems where an optimal solution is needed without any time constraints. Brute force algorithm, branch and bound algorithm and divide and conquer algorithms come under this type, where all the possible solutions are evaluated and the optimal one is selected [16, 17]. On the other hand, the approximation algorithms come as the best option after the exact algorithms for solving NP-complete problems. Greedy algorithms, simple heuristic algorithms and memetic algorithms are example of approximation algorithms. As stated earlier, the execution time increases exponentially with the size for NP-complete problems using exact algorithms, even for an ordinary size NP-complete problem it takes thousands or billions of execution time years to compute the solution using the currently available computational power. While the approximation algorithms provide an approximate result in polynomial runtime. Therefore, the approximation algorithms are preferably the better option for researcher to come up with a solution to NP-complete problems [18,19]. The proposed algorithm -min-to-min is also an approximation algorithm to solve MVC problem. We used Approximation ratio - the tool to evaluate the performance of any approximation algorithms.

This paper is extended version of our paper published [1]. In this paper we have proposed a new algorithm using the min-to-min (MtM) approach. The purpose of the proposed algorithm is to enhance the optimality and decrease the computational time. The developed approach is very simple and is based on simple heuristic method. Due to simplicity and intelligent selection of vertices in vertex cover set it save the time and improve the performance in term of optimality of the proposed algorithm.

This paper is structured as such: first the related work is discussed in detail in Section II and then the proposed algorithm is elaborated in Section III. While Section IV presents the implementation, alone with our experimental results, and discussion. Finally, Section V presents the conclusion of the paper in detail. The abbreviations with their corresponding descriptions are presented in detail in Table I.

TABLE I. ABBREVIATIONS WITH DESCRIPTIONS

Notations	Descriptions
MVC	Minimum Vertex Cover
MIS	Maximum Independent Set
MDG	Maximum Degree Greedy
VSA	Vertex Support Algorithm
MVSA	Modified Vertex Support Algorithm
MtM	Min-to-Min
AE	Absolute Error
V	Vertices
E	Edges

II. Related Work

Several approximation algorithms were proposed by various authors to deal with the problem of minimum vertex cover. Based on their performance in terms of their run time complexity, optimality, and performance on small benchmarks, a brief literature review of these popular algorithms proposed for MVC are discussed here.

The first algorithm is maximum degree greedy algorithm (MDG). The approach adopted in this algorithm is greedy. It introduces few changes in the previously existing greedy heuristic algorithm for set-cover problem by Chavatal in 1979 [20]. It uses the idea of subtracting the weight of covered vertex from all its neighbor vertices into the greedy algorithm for finding the vertex cover. Therefore, it adds all the vertices having the maximum degree to MVC. Its run time complexity for the worst case is $O(E^2)$. This algorithm fails even on small benchmarks as shown in Fig. 1(b).

On the other hand, different approaches were also used to create more efficient algorithm. Vertex Support Algorithm (VSA) is one of them.

VSA is also an approximation algorithm. The basic version of this algorithm is that it uses an adjacency matrix with binary variables 1,0; to indicate the existence of edge between two vertices with 1, and absence with 0. Its output shows if any vertex is included in vertex cover or not included. A new data structure – support of a vertex- is implemented in the algorithm. It is defined as the sum of the degree of the adjacent vertices to a given vertex of the graph. It calculates the degree and support of all the vertices, then select those vertices which have maximum support and put into vertex cover. If there are vertices with same support, then it selects the maximum degree vertex, and includes that vertex to vertex cover set [5]. The VSA has the run time complexity of $O(EV^2)$ in the worst-case scenario.

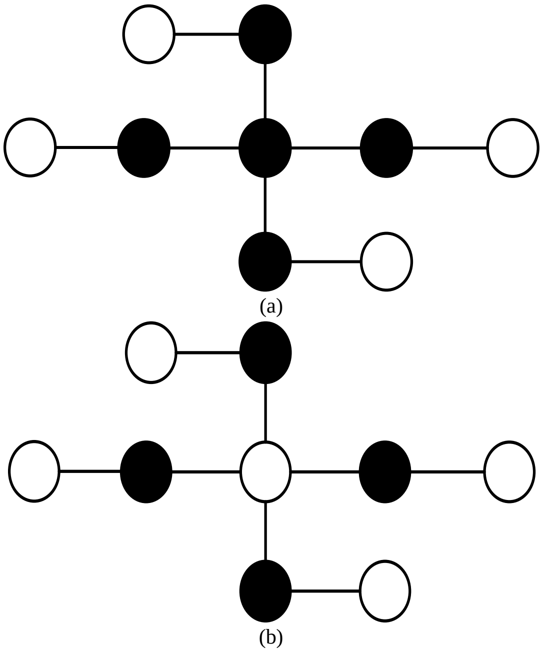


Fig. 1. (a) Optimal MVC, and (b) MVC through MDG, MVSA and VSA.

Furthermore, there is an improved and modified version of VSA named as modified vertex support algorithm (MVSA). It was proposed by Imran *et al.* in 2013[21]. The modification introduced by Imran *et al.* is in the criteria of selecting the vertex for considering it as candidate for minimum vertex cover, which, as seen above in MDG is selecting the maximum degree vertex, while in VSA the vertex which has maximum support value was selected. MVSA works in three stages. First is the analysis step, where the values of degrees and support of the vertices is calculated. Then comes the filtering process where the values are categorized as 'min_support'- vertices which have the minimum value for support and 'adj_nodes'- all the adjacent vertices to the vertices having minimum support values. And then in the last step selection is done by choosing the vertex with minimum support value, as candidate vertex for MVC. Again, the run time complexity for this algorithm is $O(EV^2 \log V)$.

Another modified version of vertex support algorithm proposed by Ahmad *et al.* [22] is known as advanced vertex support algorithm (AVSA). Ahmad *et al* introduced a tiny change in MVSA. Keeping the same data structure of support of vertex, while instead of selecting the maximum or minimum support of vertex as seen above in VSA and MVSA respectively, an alternate way was adopted by just finding all the adjacent vertex to those vertices which have minimum support. Then from all the found adjacent vertices, again selecting the one with minimum support. The computational complexity for AVSA is same as MVSA i.e. $O(EV^2 \log V)$.

Other than the above algorithms, Gujrat *et al* [23] have also contributed to the problem of minimum vertex cover by proposing the near optimal vertex cover algorithm (NOVCA). It works by successively appending the adjacent vertices to the minimum degree vertex in each step. It deals with a tie situation by choosing the neighbors of a vertex having the maximum degree instead of using neighbor of minimum degree vertex. This algorithm also has the polynomial run time complexity of $O(EV^2 \log V)$. Using NOVCA on small benchmark instances yielded failure.

In 2012, Gujral *et al* [24] further improved their algorithm and came up with a modified version of NOVCA i.e NOVCA-II. NOVCA-II is simply the incorporation of idea that a vertex cover should be built by adding the vertex in decreasing order of their degree and in the situation of tie, it uses the converse of NOVCA i.e. it chooses the vertices which have the minimum sum of degrees of their neighbor vertex, instead of selecting the neighbors of vertex having maximum degree as in NOVCA. By using few graph instances, it became evident that both algorithms of Gujral *et al* also fail on small benchmarks.

More investigation has been done and Li *et al* in 2011[25] came up with a new idea of max-share of degree heuristic algorithm (Max-I). This algorithm works based on the concept that, only those vertices should be added to minimum vertex cover, which helps to reduce the degree of its neighboring vertices as much as possible, or ultimately if zero. It follows random selection in cases of tie. Max-I has the worst run time complexity of $O(n^3)$.

However, in 2014, Imran *et al* [26] again came up with his second method for MVC called Degree Contribution Algorithm (DCA). A new data structure -Degree contribution- was introduce. Degree contribution is used for graph processing parameters. It helps in taking the complete graph to determine the values for each vertex. Degree contribution of a vertex is the total number of degrees of the vertex and sum of all the vertices with same degree in the whole graph. The degree contribution for each vertex determines either it is efficient to choose that vertex for MVC or not. So, DCA works by calculating the degree of every vertex along with its degree contribution value**.** Then only those vertices are added to minimum vertex cover whose degree contribution value is higher than others. After adding a vertex with higher degree contribution value, all its neighbor edges are removed and the same steps are repeated for the remaining vertices, until no edge was left in the graph. The execution time of DCA is $O(VE)$. A memetic algorithm was also proposed by Jovanovic *et al* in 2011 [27].

Based on heuristic nature of real ants searching for food, ant colony algorithm was proposed for minimum weight vertex cover problem (MWVC). As memetic framework of ant colony, bee colony and bat colony were mostly used for optimization problems. However, the ant colony algorithm was initially applied to the traveling salesman problems (TSP), where a fully connected graph with weighted edges were considered. While in case of MVC it is not necessary for the graph to be fully connected and un-weighted. So, Jovanovic *et al*. used arbitrary edges to make the given graph a connected graph and gave weight to each vertex to make it complacent for the algorithm. Hence, converting the graph from MVC to TSP and applying the ant colony algorithm was the main aim. For un-weighted graphs, this algorithm is much harder to apply. Lastly an unusual attempt in proposing a unique algorithm for MVC was undertaken by Halldórsson *et al* in [28]. Based on the greedy strategy the greed is good (GGA) algorithm was proposed. This algorithm tackled the maximum independent set (MIS) problem. The "Network Bench Model" was used in this algorithm. The degree for every vertex is found then only the minimum degree vertex is included in MIS. Once, MIS is found then MVC can be easily fond by taking complement of MIS from the set of all vertices.

Li et al. [29] proposed a new local search technique to solve the minimum vertex cover problem called NuMVC. NuMVC comprised of three main stages. In the first stage the introduction of four rules is carried out. In the second stage a technique is introduced called configuration checking in order to reduce cycling in local search. In the third phase a method is introduced to reduce the searching time in order to improve the optimality and reduce the computation time. In [30] two new extensions of the conventional vertex cover problem are introduced. The detail presentation of the two new methods is carried out in detail.

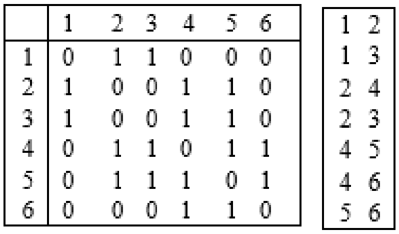
In the literature review section different algorithms are discussed in detail from different aspects with the aim to find the weakness and strengths of the existing algorithms for minimum vertex cover problem. These algorithms have weakness in one way or another way, some are fast but fail to provide optimal results on many cases, some algorithms provide good optimality results but are very slow. The proposed method is introduced to tackle all these issues.

III. Proposed Algorithm

The algorithm put forth here, has a different and unique approach than any algorithms found in the literature review. The min-to-min (MtM) consists of three basic steps of working through an undirected and unweighted graph with any number of vertices. The initial step in the algorithm is to find the degree of every vertices in the provided graph, then comes the next step of finding the minimum degree vertex and getting its adjacent vertices. If there are more vertices with same minimum degree than we will choose the first vertex with minimum degree. After getting the minimum degree vertex and all of its adjacent vertices. The last step is to again find the minimum degree vertex from the adjacent vertices of already found minimum degree vertex, and this vertex is considered as a candidate for MVC. Once the candidate vertex is selected from the adjacent vertices, all its edges are deleted, and the vertex is appended to MVC. This process is repeated till no edge is left in the provided graph.

A. Data Structure

In our work, we implemented the edge list data structure for representing graphs in computer memory leveraging the properties of saving computation time for sparse graphs using list indexing to access any edge and also fully exploiting the advantages of saving space as well instead of using adjacency list for the graph representation. In case of the space usage the performance of edge list is $O(x + y\ )$ for the representation of the graph having x edges and y vertices. And in case of the run time, edge list performs well in counting the number of vertices or edges and looping an iteration over the edges or vertices [31]. The iterations over the edge list in our case using the method find () has a runs time of $O(1)$. While the method min(array) iterates through the required array in run time of $O(y)$, y being the total vertices. Furthermore, removing any edges associated with a vertex has the complexity of run time $O(x)$, which reduces exponentially as by each iteration our graph gets smaller with few vertices and associated edges. On the other hand, adjacency matrix- which is mostly used as data structure for graph representation- has both performance drawbacks of run time as well as space storage problems for our purpose. For instance, the $O(x^2)$ space usage is the worse than the $O(x + y)$ space usage by edge list, moreover most real-world graphs are sparse which makes it more disadvantageous to implement adjacency matrix as depicted in Fig. 2(a) and (b) the adjacency matrix and edge list representation respectively of the graph from Fig. 1(a).



	1	2	3	4	5	6
1	0	1	1	0	0	0
2	1	0	0	1	1	0
3	1	0	0	1	1	0
4	0	1	1	0	1	1
5	0	1	1	1	0	1
6	0	0	0	1	1	0

(a)

1	2
1	3
2	4
2	3
4	5
4	6
5	6

(b)

Fig. 2. Ardency Matrix and Edge List.

B. Terminologies

Following are some terminologies that we have used in the proposed pseudo code as illustrated in Fig. 3 deg (): used for degree calculation, **| |:** used for absolute values, min(): for minimum degree calculation in the graph, adj (): to calculate the neighbors/adjacent vertices of a node/vertex and C is used for cover.

Pseudo code for Clever Steady Strategy Algorithm

```
Start:
Input: = G (V, E)
Output: = C
        FOR i← 1 to n {
            $d_i \leftarrow deg(V_i)$
        }
        While G ≠ θ {
            $M_i \leftarrow min(d_i)$
            IF $|M_i| = 1$ {
                $N_i \leftarrow adj\ (M_i)$
            }
            IF $|M_i| = 1$ {
                $N_i \leftarrow adj\ (M_i)$
                $MN_i \leftarrow adj\ (M_i)$
                $C \leftarrow adj\ (M_i)$
            }
        }
                Display C
  Stop
```

Fig. 3. Pseudo Code of the Min-to-Min Algorithm.

C. Flow Chart of the Algorithm

The flow chart of the proposed algorithm is provided in Fig. 4. In the flow diagram we have used the same terminologies as given in the pseudo code of the proposed MtM algorithm.

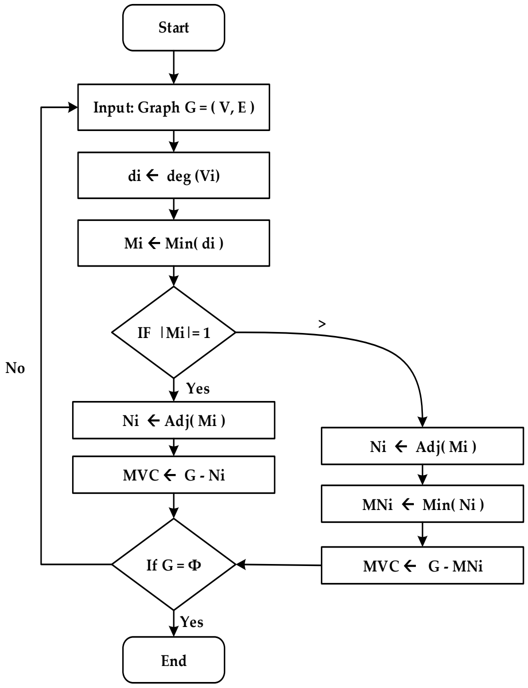


Fig. 4. Flow Chart of the Proposed Algorithm.

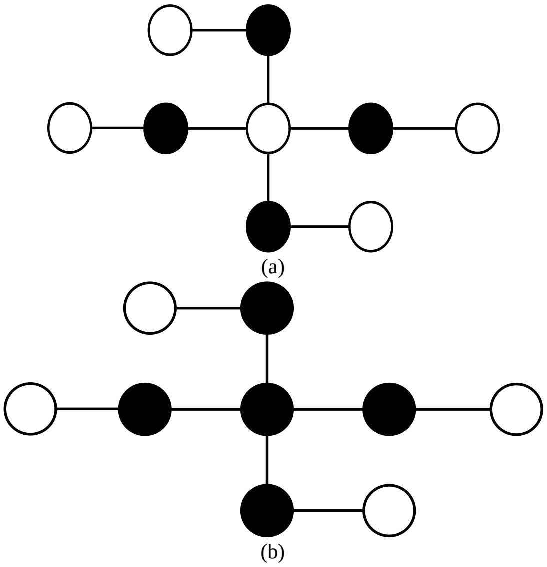
IV. Implementation, Empirical Results and Discussion

A. Implementation Setup

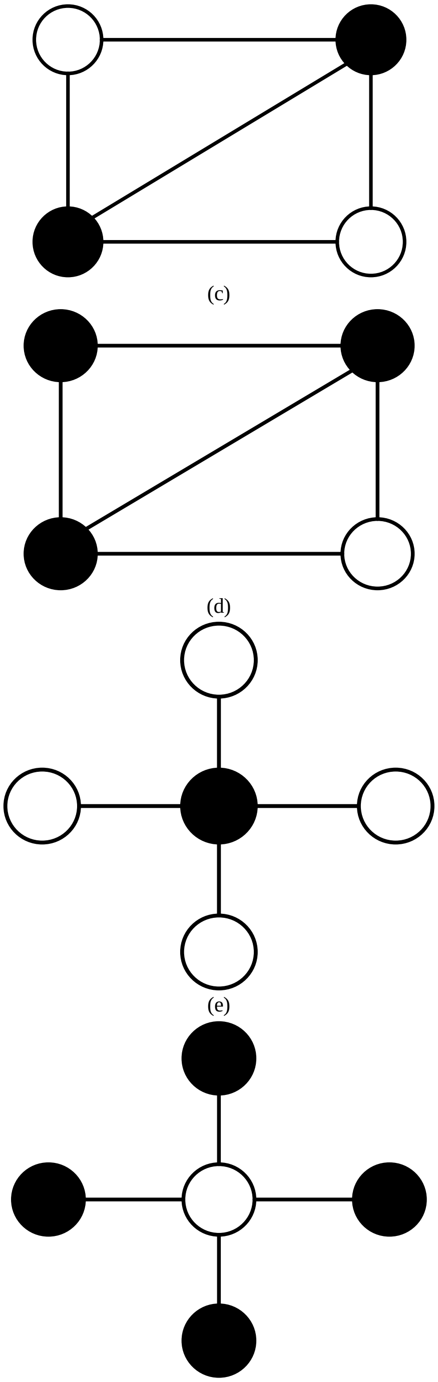
Here, the section deals with the empirical results gained through implementation of MtM on various benchmark instances of up-to-date popular libraries. By cross evaluating the results of MtM to different well-known proposed algorithms, as discussed in literature review also, the tables were created. And by the run time complexity and optimality the result of MtM were compare with VSA, MDG and MVSA in the tabular form. The coding of the MtM algorithm has been done in MATLAB R2014a version 7.10.0.499 on an Intel core i5 system having windows 7 operating system.

B. Results on Small Instances

Implementing the proposed algorithm on small benchmark revealed an interesting result and optimal performance of MtM as compare to performance of other algorithms discussed in the literature review. As shown in Fig. 5(a) the MVC by MtM for the given graph is 4 vertices which is optimal and the approximation ratio ρi = 1. Which is also optimal. While in Fig. 5(b) the same graph is covered by 5 vertices according to MDG and ρi = 1.250. while the given graph in Fig. 5(c) is covered using just 2 vertices by MtM, which is optimal solution and ρi = 1, while it can be seen in Fig. 5(d) the same graph is covered by VSA using 3 vertices with ρi of 1.5. Moreover, most astonishing performance of MtM appears in Fig. 5(e) which is covered by just 1 vertex- which is the optimal solution. But the same graph shown in Fig. 5(f) is covered using 4 vertices by MVSA with ρi = 4. As the solution by MtM for the small benchmark instances is optimal and it outperformed all other existing algorithms, hence the MtM is much efficient as well as fast in relation to the already proposed algorithms in literature.


(a)

(b)


(c)

(d)

(e)

(f)

Fig. 5. (a) MtM ρi = 1, (b) MDG ρi = 1.25, (c) MtM ρi = 1, (d) VSA ρi = 1.5, (e) MtM ρi = 1, (f) MVSA ρi = 4.

C. Perfomance of MtM on Large Benchmark Instances

Similarly, the performance by MtM was evaluated on large benchmarks instance along with the comparison of its performance with other well-known existing algorithms. The used libraries for the large benchmark instances were DIMACS and BHOSHLIB. In Table II the performance of MtM is evaluated against VSA, MDG and MVSA based on the total vertices covered by the proposed algorithm along with the counterpart algorithms. The approximation ratio of each algorithm for each benchmark instance as well as the absolute error was evaluated in Table III. In Table VI the worst-case approximation ratios and average cases approximation ratios have been calculated for MtM, VSA, MDG and MVSA for MVC. The approximation ratios of MtM for both cases are much smaller in comparison with the other state of art algorithms which indicates that the working of MtM algorithm is much efficient than the counterpart algorithms in literature. Similarly, the approximation ratios for the worst and average case have been calculated for MtM, MDG, VSA, and MVSA for maximum independent set problem (MIS) in Table VII. The approximation ratios also indicate that the results from the proposed method are efficient as compared to the other algorithms for MIS problem.

TABLE II. PERFORMANCE OF MTM, MDG, VSA AND MVSA BENCHMARK INSTANCES FOR MINIMUM VERTEX COVER (MVC)

S. NO	Benchmarks	V	C*	MtM	MDG	VSA	MVSA
				|MVC|	|MVC|	|MVC|	|MVC|
1	graph50_6	50	38	38	38	44	38
2	graph50_10	50	35	35	35	41	35
3	graph100_1	100	60	60	60	95	60
4	graph100_10	100	70	70	70	96	70
5	graph200_5	200	150	150	150	184	150
6	graph500_1	500	350	350	350	485	350
7	graph500_2	500	400	400	400	484	400
8	graph500_5	500	290	290	290	454	290
9	phat300-1	300	292	293	293	292	294
10	phat300-2	300	275	275	278	275	279
11	phat300-3	300	264	266	269	264	272
12	phat700-1	700	689	692	693	689	692
13	phat700-2	700	656	658	660	656	660
14	phat700-3	700	638	640	642	638	649
15	johnson8-2-4	28	24	24	24	24	24
16	johnson8-4-4	70	56	56	62	56	56
17	johnson16-2-4	120	112	112	112	112	112
18	johnson32-2-4	496	480	480	480	480	480
19	sanr200_0.7	200	182	183	184	182	186
20	sanr200_0.9	200	158	164	164	158	163
21	sanr400_0.5	400	387	388	392	387	389
22	sanr400_0.7	400	379	382	384	379	381
23	fbr35-17-2	595	560	565	570	573	424
24	fbr_30_15_5	450	420	426	429	429	565
25	c125	125	91	94	93	91	95
25	C250.9	250	206	211	211	206	211
27	C500.9	500	443	448	453	443	449
28	C2000.9	2000	1922	1927	1944	1923	1937
29	brock200_1	200	188	190	190	188	191
30	brock200_4	200	183	185	192	183	193
31	gen200_p0.9_44	200	156	164	165	156	166
32	hamming6-2	64	32	32	32	32	32
33	hamming6-4	64	60	60	60	60	60
34	hamming8-2	256	128	128	128	128	128
35	hamming8-4	256	240	240	240	240	240
36	hamming10-2	1024	512	512	512	512	512
37	dsjc-500	500	487	489	491	487	489
38	killer4	171	160	160	164	160	160
39	killer5	776	749	754	764	749	754
40	c-fat200-1	200	188	188	188	188	188
41	c-fat200-2	200	176	176	176	176	176
42	c-fat200-5	200	142	142	142	144	142
43	c-fat500-1	500	486	486	486	486	486
44	c-fat500-2	500	474	474	474	474	474
45	c-fat500-5	500	436	436	436	436	436
46	c-fat500-10	500	374	374	374	374	374
47	MANN_a27.clq.b	378	252	253	261	253	253

TABLE III. PERFORMANCE AND COMPARISON OF MtM, AND MDG, VSA, MVSA USING APPROXIMATION RATIO AND ABSOLUTE ERROR FOR MVC

S. NO	Approximation Ratio				Absolute Error			
	MtM	MDG	VSA	MVSA	MtM	MDG	VSA	MVSA
1	1.0000	1.0000	1.1579	0.8636	0	0	6	0
2	1.0000	1.0000	1.1714	0.8537	0	0	6	0
3	1.0000	1.0000	1.5833	0.6316	0	0	35	0
4	1.0000	1.0000	1.3714	0.7292	0	0	26	0
5	1.0000	1.0000	1.2267	0.8152	0	0	34	0
6	1.0000	1.0000	1.3857	0.7216	0	0	135	0
7	1.0000	1.0000	1.2100	0.8264	0	0	84	0
8	1.0000	1.0000	1.5655	0.6388	0	0	164	0
9	1.0034	1.0000	0.9966	1.0068	1	1	0	2
10	1.0000	1.0109	0.9892	1.0145	0	3	0	4
11	1.0076	1.0113	0.9814	1.0303	2	5	0	8
12	1.0044	1.0014	0.9942	1.0044	3	4	0	3
13	1.0030	1.0030	0.9939	1.0061	2	4	0	4
14	1.0031	1.0031	0.9938	1.0172	2	4	0	11
15	1.0000	1.0000	1.0000	1.0000	0	0	0	0
16	1.0000	1.1071	0.9032	1.0000	0	6	0	0
17	1.0000	1.0000	1.0000	1.0000	0	0	0	0
18	1.0000	1.0000	1.0000	1.0000	0	0	0	0
19	1.0055	1.0055	0.9891	1.0220	1	2	0	4
20	1.0380	1.0000	0.9634	1.0316	6	6	0	5
21	1.0026	1.0103	0.9872	1.0052	1	5	0	2
22	1.0079	1.0052	0.9870	1.0053	3	5	0	2
23	1.0089	1.0088	1.0053	0.7400	5	10	13	136
24	1.0143	1.0070	1.0000	1.3170	6	9	9	145
25	1.0330	0.9894	0.9785	1.0440	3	2	0	4
25	1.0243	1.0000	0.9763	1.0243	5	5	0	5
27	1.0113	1.0112	0.9779	1.0135	5	10	0	6
28	1.0026	1.0088	0.9892	1.0073	5	22	1	15
29	1.0106	1.0000	0.9895	1.0160	2	2	0	3
30	1.0109	1.0378	0.9531	1.0546	2	9	0	10
31	1.0513	1.0061	0.9455	1.0641	8	9	0	10
32	1.0000	1.0000	1.0000	1.0000	0	0	0	0
33	1.0000	1.0000	1.0000	1.0000	0	0	0	0
34	1.0000	1.0000	1.0000	1.0000	0	0	0	0
35	1.0000	1.0000	1.0000	1.0000	0	0	0	0
36	1.0000	1.0000	1.0000	1.0000	0	0	0	0
37	1.0041	1.0041	0.9919	1.0041	2	4	0	2
38	1.0000	1.0250	0.9756	1.0000	0	4	0	0
39	1.0067	1.0133	0.9804	1.0067	5	15	0	5
40	1.0000	1.0000	1.0000	1.0000	0	0	0	0
41	1.0000	1.0000	1.0000	1.0000	0	0	0	0
42	1.0000	1.0000	1.0141	0.9861	0	0	2	0
43	1.0000	1.0000	1.0000	1.0000	0	0	0	0
44	1.0000	1.0000	1.0000	1.0000	0	0	0	0
45	1.0000	1.0000	1.0000	1.0000	0	0	0	0
46	1.0000	1.0000	1.0000	1.0000	0	0	0	0
47	1.0040	1.0316	0.9693	1.0000	1	9	1	1

The Absolute error is calculated for each algorithm on each benchmark instance, in Table III. Similarly, the MtM is also compared with other algorithms in terms of getting the Maximum independent set in Table IV. And the approximation ratio along with Absolute error for MIS of each algorithm is presented in Table V.

Mathematically Absolute Error is defined as, the magnitude of the difference between the measured/ obtained or approximated value of a quantity $x_{approx.}$ and the actual or optimal value $x_{optimal}$.

$$\Delta X = \left| x_{optimal} - x_{approx} \right| \qquad (1)$$

In our case, absolute error ΔX served as the best indicator of the performance of the algorithm. Like the value ΔX shows how close the result of the algorithm is to the actual solution. The smaller the value of absolute error ΔX, the better the solution can be considered. Large value of ΔX indicate worst performance and a poor solution. Table III shows the approximation ratio, along with Absolute error values, which are calculated for the measurement of performance with respect to the optimal output of each algorithm.

TABLE IV. PERFORMANCE OF MTM, MDG, VSA AND MVSA BENCHMARK INSTANCES FOR MAXIMUM INDEPENDENT SET (MIS)

S. NO	Benchmarks	V	C*	MtM	MDG	VSA	MVSA
				|MIS|	|MIS|	|MIS|	|MIS|
1	graph50_6	50	12	12	12	6	12
2	graph50_10	50	15	15	15	9	15
3	graph100_1	100	40	40	40	5	40
4	graph100_10	100	30	30	30	4	30
5	graph200_5	200	50	50	50	16	50
6	graph500_1	500	150	150	150	15	150
7	graph500_2	500	100	100	100	16	100
8	graph500_5	500	210	210	210	46	210
9	phat300-1	300	8	7	7	8	6
10	phat300-2	300	25	25	22	25	21
11	phat300-3	300	36	34	31	36	28
12	phat700-1	700	11	8	7	11	8
13	phat700-2	700	44	42	40	44	40
14	phat700-3	700	62	60	58	62	51
15	johnson8-2-4	28	4	4	4	4	4
16	johnson8-4-4	70	14	14	8	14	14
17	johnson16-2-4	120	8	8	8	8	8
18	johnson32-2-4	496	16	16	16	16	16
19	sanr200_0.7	200	18	17	16	18	14
20	sanr200_0.9	200	42	36	36	42	37
21	sanr400_0.5	400	13	12	8	13	11
22	sanr400_0.7	400	21	18	16	21	19
23	fbr35-17-2	595	35	30	25	22	171
24	fbr_30_15_5	450	30	24	21	21	-115
25	c125	125	34	31	32	34	30
25	C250.9	250	44	39	39	44	39
27	C500.9	500	57	52	47	57	51
28	C2000.9	2000	78	73	56	77	63
29	brock200_1	200	12	10	10	12	9
30	brock200_4	200	17	15	8	17	7
31	gen200_p0.9_44	200	44	36	35	44	34
32	hamming6-2	64	32	32	32	32	32
33	hamming6-4	64	4	4	4	4	4
34	hamming8-2	256	128	128	128	128	128
35	hamming8-4	256	16	16	16	16	16
36	hamming10-2	1024	512	512	512	512	512
37	dsjc-500	500	13	11	9	13	11
38	killer4	171	11	11	7	11	11
39	killer5	776	27	22	12	27	22
40	c-fat200-1	200	12	12	12	12	12
41	c-fat200-2	200	24	24	24	24	24
42	c-fat200-5	200	58	58	58	56	58
43	c-fat500-1	500	14	14	14	14	14
44	c-fat500-2	500	26	26	26	26	26
45	c-fat500-5	500	64	64	64	64	64
46	c-fat500-10	500	126	126	126	126	126
47	MANN_a27.clq.b	378	126	125	117	125	125

Approximation ratio is the ratio of number of vertices in MCV found by an algorithm to the optimal solution of the MVC. Mathematically, approximation ratio is defined- as in Eq. (2).

$$\rho_{i} = \frac{A_i}{OPT_i} \qquad (2)$$

In the above equation, A denotes the approximate result and *i* represents the number of instances, OPT*i* represents the optimal result.

For MVC ρi = 1 is the optimal solution while for the existing algorithms ρi ≥ 1 always. A ratio very close to 1 means the result found is better and almost near to optimal solution while the more the ratio deviates from 1, the solution gets poor and worst. On the other hand, for MIS ρi = 1 also means the result provided by the algorithm is optimal but all existing algorithms will always have ρi ≤ 1, as for MVC for MIS if the deviation from 1 is very small then the solution is considered as the near to optimal, but a deviation of 0.1 means the solution is not better one while high values of the deviation indicates the solution is the worst. Thus, for MIS approximation ratio ≤1 and for MVC approximation ratio ≥ 1.

TABLE V. PERFORMANCE OF MTM, MDG, VSA AND MVSA BASED ON APPROXIMATION RATIO AND ABSOLUTE ERROR FOR MIS

S. NO	**Approximation Ratio**				**Absolute Error**			
	MtM	**MDG**	**VSA**	**MVSA**	**MtM**	**MDG**	**VSA**	**MVSA**
1	1.0000	1.0000	0.5000	1.0000	0	0	6	0
2	1.0000	1.0000	0.6000	1.0000	0	0	6	0
3	1.0000	1.0000	0.1250	1.0000	0	0	35	0
4	1.0000	1.0000	0.1333	1.0000	0	0	26	0
5	1.0000	1.0000	0.3200	1.0000	0	0	34	0
6	1.0000	1.0000	0.1000	1.0000	0	0	135	0
7	1.0000	1.0000	0.1600	1.0000	0	0	84	0
8	1.0000	1.0000	0.2190	1.0000	0	0	164	0
9	0.8750	0.8750	1.0000	0.7500	1	1	0	2
10	1.0000	0.8800	1.0000	0.8400	0	3	0	4
11	0.9444	0.8611	1.0000	0.7778	2	5	0	8
12	0.7273	0.6364	1.0000	0.7273	3	4	0	3
13	0.9545	0.9091	1.0000	0.9091	2	4	0	4
14	0.9677	0.9355	1.0000	0.8226	2	4	0	11
15	1.0000	1.0000	1.0000	1.0000	0	0	0	0
16	1.0000	0.5714	1.0000	1.0000	0	6	0	0
17	1.0000	1.0000	1.0000	1.0000	0	0	0	0
18	1.0000	1.0000	1.0000	1.0000	0	0	0	0
19	0.9444	0.8889	1.0000	0.7778	1	2	0	4
20	0.8571	0.8571	1.0000	0.8810	6	6	0	5
21	0.9231	0.6154	1.0000	0.8462	1	5	0	2
22	0.8571	0.7619	1.0000	0.9048	3	5	0	2
23	0.8571	0.7143	0.6286	4.8857	5	10	13	136
24	0.8000	0.7000	0.7000	-3.8333	6	9	9	145
25	0.9118	0.9412	1.0000	0.8824	3	2	0	4
25	0.8864	0.8864	1.0000	0.8864	5	5	0	5
27	0.9123	0.8246	1.0000	0.8947	5	10	0	6
28	0.9359	0.7179	0.9872	0.8077	5	22	1	15
29	0.8333	0.8333	1.0000	0.7500	2	2	0	3
30	0.8824	0.4706	1.0000	0.4118	2	9	0	10
31	0.8182	0.7955	1.0000	0.7727	8	9	0	10
32	1.0000	1.0000	1.0000	1.0000	0	0	0	0
33	1.0000	1.0000	1.0000	1.0000	0	0	0	0
34	1.0000	1.0000	1.0000	1.0000	0	0	0	0
35	1.0000	1.0000	1.0000	1.0000	0	0	0	0
36	1.0000	1.0000	1.0000	1.0000	0	0	0	0
37	0.8462	0.6923	1.0000	0.8462	2	4	0	2
38	1.0000	0.6364	1.0000	1.0000	0	4	0	0
39	0.8148	0.4444	1.0000	0.8148	5	15	0	5
40	1.0000	1.0000	1.0000	1.0000	0	0	0	0
41	1.0000	1.0000	1.0000	1.0000	0	0	0	0
42	1.0000	1.0000	0.9655	1.0000	0	0	2	0
43	1.0000	1.0000	1.0000	1.0000	0	0	0	0
44	1.0000	1.0000	1.0000	1.0000	0	0	0	0
45	1.0000	1.0000	1.0000	1.0000	0	0	0	0
46	1.0000	1.0000	1.0000	1.0000	0	0	0	0
47	0.9921	0.9286	0.9921	0.9921	1	9	1	1

TABLE VI. APPROXIMATION RATIOS FOR WORST AND AVERAGE CASE FOR MTM, MDG, MVSA, AND VSA FOR MVC

Algorithms	**Worst case for MIS** **ρ_i**	**Average for MIS** **ρ_i**
MtM	0.7272	0.9476
MDG	0.4444	0.8804
VSA	0.1000	0.8602
MVSA	-3.8333	0.9010

TABLE VII. APPROXIMATION RATIOS FOR WORST AND AVERAGE CASE FOR MDG, MVSA AND VSA FOR MIS

Algorithms	**Worst case for MVC** **ρ_i**	**Average for MCV** **ρ_i**
MtM	1.0512	1.0054
MDG	1.1071	1.0065
VSA	1.5833	1.0467
MVSA	1.3170	0.9681

V. Conclusion and Future Work

This paper proposes a new algorithm – Min to Min (MtM) to tackle the NP- hard optimization problems of MVC as well as MIS problems of graph theory. The algorithm was well tested first on small benchmark instances as well as then it was investigated on benchmarks instance of libraries like DIMACS and BHOSLIB. The obtained experimental results were cross compared with other extant algorithms of literature. After the obtained results it was proved that MtM outperformed algorithms like MDG, VSA and MVSA on the parameters of runtime performance as well as approximation ratio and absolute error. The results are well demonstrated using tables and visual graph diagrams. Hence, the proposed algorithm is an efficient way to solve real world problems of MVC and MIS in different areas of application. It is simple and fast to implement, and it gives near to optimal solutions on the grounds of run time, storage and parameter of approximation ratio, and the absolute error. While it outperforms other existing algorithms in worst and average cases.

In the future we would like to add some more tweaks to the proposed algorithm to improve it in term of optimality and apply on other NP-complete problems as such as graph coloring, maximum clique etc.

Acknowledgment

The authors would like to acknowledge Professor Neal Gallagher who motivated and encouraged us for research at the University of Central Asia. Further, the authors would like to thank Sumaino Shakarbekova for proofreading the paper.

References

[1] M. Fayaz, and S. Arshad. "Clever Steady Strategy Algorithm: A Simple and Efficient Approximation Algorithm for Minimum Vertex Cover Problem." 2015 13th International Conference on Frontiers of Information Technology (FIT). IEEE, 2015.

[2] D. B. West, Introduction to graph theory. Prentice hall Upper Saddle River, NJ, 1996.

[3] V. Kann, "On the approximability of NP-complete optimization problems," Royal Institute of Technology Stockholm, 1992.

[4] F. N. Abu-Khzam, M. A. Langston, P. Shanbhag, and C. T. J. A. Symons, "Scalable parallel algorithms for FPT problems," vol. 45, no. 3, pp. 269-284, 2006.

[5] S. Balaji, V. Swaminathan, K. J. W. A. o. S. Kannan, Engineering, and Technology, "Optimization of unweighted minimum vertex cover," vol. 43, pp. 716-729, 2010.

[6] S. Richter, M. Helmert, and C. Gretton, "A stochastic local search approach to vertex cover," in Annual Conference on Artificial Intelligence, 2007, pp. 412-426: Springer.

[7] B. Verweij and K. Aardal, "An optimisation algorithm for maximum independent set with applications in map labelling," in European Symposium on Algorithms, 1999, pp. 426-437: Springer.

[8] M. Weigt and A. K. J. P. r. l. Hartmann, "Number of guards needed by a museum: A phase transition in vertex covering of random graphs," vol. 84, no. 26, p. 6118, 2000.

[9] S. A. Cook, "The complexity of theorem-proving procedures," in Proceedings of the third annual ACM symposium on Theory of computing, 1971, pp. 151-158.

[10] R. M. Karp, "Reducibility among combinatorial problems," in Complexity of computer computations: Springer, 1972, pp. 85-103.

[11] M. Xiao and H. J. D. A. M. Nagamochi, "An exact algorithm for maximum independent set in degree-5 graphs," vol. 199, pp. 137-155, 2016.

[12] D. Kagaris and S. Tragoudas, "Maximum independent sets on transitive graphs and their applications in testing and CAD," in International Conference on Computer Aided Design: Proceedings of the 1997 IEEE/ACM international conference on Computer-aided design, 1997, vol. 9, no. 13, pp. 736-740.

[13] T. A. Feo, M. G. Resende, and S. H. J. O. R. Smith, "A greedy randomized adaptive search procedure for maximum independent set," vol. 42, no. 5, pp. 860-878, 1994.

[14] M. Garey and D. Johnson, "Computers and Intractability: A guide to the Theory of NP-Completeness, WH Freedman, San Francisco," 1979.

[15] M. R. Garey, D. S. J. C. Johnson, and Intractability, "A Guide to the Theory of NP-Completeness," pp. 37-79, 1990.

[16] E. Balas and J. J. A. Xue, "Weighted and unweighted maximum clique algorithms with upper bounds from fractional coloring," vol. 15, no. 5, pp. 397-412, 1996.

[17] R. Carraghan and P. M. J. O. R. L. Pardalos, "An exact algorithm for the maximum clique problem," vol. 9, no. 6, pp. 375-382, 1990.

[18] L. A. Hall and D. S. Hochbaum, "Approximation lgorithms for NP-hard problems, chapter approximation algorithms for scheduling," ed: PWS Publishing Company, Boston, MA, 1996.

[19] G. Karakostas, "A better approximation ratio for the vertex cover problem," in International Colloquium on Automata, Languages, and Programming, 2005, pp. 1043-1050: Springer.

[20] V. J. M. o. o. r. Chvatal, "A greedy heuristic for the set-

[21] covering problem," vol. 4, no. 3, pp. 233-235, 1979.

[22] K. Imran, K. J. R. J. o. C. Hasham, and I. T. S.

[23] "Modified Vertex Support Algorithm: A New approach for approximation of Minimum vertex cover," vol. 2320, p. 6527, 2013.

[24] I. Ahmad, M. J. I. J. o. A. S. Khan, and Technology, "AVSA, modified vertex support algorithm for approximation of MVC," vol. 67, pp. 71-78, 2014.

[25] S. Gajurel and R. J. P. C. S. Bielefeld, "A Simple NOVCA: Near Optimal Vertex Cover Algorithm," vol. 9, pp. 747-753, 2012.

[26] S. Gajurel and R. J. I. J. o. E. A. Bielefeld, "A fast near optimal vertex cover algorithm (novca)," vol. 3, pp. 9-18, 2012.

[27] S. Li, J. Wang, J. Chen, and Z. J. J. Wang, "An Algorithm for Minimum Vertex Cover Based on Max-I Share Degree," vol. 6, no. 8, pp. 1781-1788, 2011.

[28] I. K. a. H. Khan, "Degree Contribution Algorithm for Approximation of MVC," International Journal of Hybrid Information Technology, vol. 7, no. 5, pp. 183-190, 14 October 2014 2014.

[29] R. Jovanovic and M. J. A. S. C. Tuba, "An ant colony optimization algorithm with improved pheromone correction strategy for the minimum weight vertex cover problem," vol. 11, no. 8, pp. 5360-5366, 2011.

[30] M. M. H. a. J. Radhakrishnan, "Greed is Good: Approximating Independent Sets in Sparse and Bounded-Degree Graphs," Algorithmica, vol. 18, pp. 145-163, 1997. Springer-Verlag New York Inc.

[31] Li, R., Hu, S., Cai, S., Gao, J., Wang, Y., & Yin, M. (2020). NuMWVC: A novel local search for minimum weighted vertex cover problem. Journal of the Operational Research Society, 71(9), 1498-1509.

[32] Akrida, E. C., Mertzios, G. B., Spirakis, P. G., & Zamaraev, V. (2020). Temporal vertex cover with a sliding time window. Journal of Computer and System Sciences, 107, 108-123.

[33] R. T. Michael T. Goodrich, Michael H. Goldwasser, "Graph Algorithms," in Data Structures & Algorithms in Python: John Wiley & Sons, Inc., 2013, pp. 628-633.

Developing an Intelligent Framework for Improving the Quality of Service in the Government Organizations in the Kingdom of Saudi Arabia

Abdulelah Abdallah AlGosaibi[1], Abdul Rahaman Wahab Sait[2], Abdulaziz Fahad AlOthman[3], Shadan AlHamed[4]
College of Computer Science and Information Technology, King Faisal University, Al-Ahsa, Kingdom of Saudi Arabia[1]
Department of Documents and Archives, King Faisal University, Al-Ahsa, Kingdom of Saudi Arabia[2, 3, 4]

***Abstract*—The Kingdom of Saudi Arabia is enhancing the services and applications in government organizations through the number of systems that generate a massive amount of data through Big Data technology. Recently, the Global Artificial Intelligent Summit 2020, Saudi Data and Artificial Intelligence Authority (SDAIA), NEOM have launched an Artificial Intelligence (AI) strategy that aligns with the Kingdom Vision 2030. AI opens a wide door for opportunities and new strategies that will narrow the gap in the skillset of individuals and promote research and innovation in the IT industry. Organizations lack advanced techniques to evaluate the performance of individuals and departments that supports improving the quality of service. The introduction of AI-based applications in the government and private sectors will facilitate decision-makers in tracking and optimizing the efficiency of departments and individuals. This research aims to develop an intelligent framework for government organizations to improve the quality of services rendered to customers and businesses. In addition, it highlights the importance of AI policies in archiving metadata. This paper presents a framework for an organization that contains Chatbot, Sentiment Analysis, and Key Performance Indicators to improve the services. A synthetic dataset is employed as a testbed to evaluate the performance of the framework. The outcome of this study shows that the proposed framework able to improve the performance of organizations. Using this proposed framework, organizations can build a mechanism for their workforce to retrieve meaningful information. Moreover, it provides significant features include efficient data extraction, data management, and AI-based security for effective document management.**

Keywords—Key performance indicators; big data; hierarchical analysis; artificial intelligence; privacy policies; metadata; pattern generation

I. Introduction

The vision 2030 of the Kingdom of Saudi Arabia has initiated many plans and programs to enhance the performance of Education, Medicine, and other core sectors. Skill development and performance improvement are the two essential goals of vision 2030 [1]. In recent days, data produced widely from various resources. Artificial Intelligence (AI) and Big Data (BD) are the key players in transforming and empowering modern organizations process that improves decision-making, trend analysis, forecasts, etc. [2][3]. Valuable information from them by applying BD analytics to stored data. And, sophisticated AI algorithms are used to obtain insights into consumer behaviour in order to grasp current patterns and stand high among competitors [4]. BD monitors the performance of workers, design precise delivery models, predictive maintenance, etc. Few organizations use this technology to create a smart workplace and track their employees [5]. The use of metadata in global enterprises is vital in determining the success of a sustainable organization and in gathering insight. Metadata provides meta-information about communication that occurred between two or more individuals or departments. Identifying the invaluable management strategies that can minimize costs, assess how far a company advances and equate them to its organizational objectives [6][7]. This inspiration from top-level motivates organizations to enhance current functionally and lunch new advanced one. Therefore, there is a demand for intelligent data analytical tools for measuring the performance of each department and individual in Government Organization (GO). The availability of sophisticated tools and techniques provide an opportunity for decision-maker to make an effective decision that improves individual performance which leads to organization success [6][9][10]. Data relating to document flow in an organization would be a supporting and influencing factor for the 2030 vision and objective. It can be used to monitor the performance of administrative departments and employees. The data collection process will begin with the creation of correspondence in a particular department and collects data, including receiver information, degree of confidentiality, urgency, and automatically created data, such as the name of the sender, correspondence time, and data, and transmission information. Such data can be employed to provide a base for achieving an individual's efficiency. Recently, in the Global Summit 2020, the strategy for AI and data is announced by Saudi Data and Artificial Intelligence Authority (SDAIA). The plan supports individual to learn AI technology and develop their skills. Also, it assists organizations in applying AI to utilize the resources effectively. Furthermore, GOs are following the rules and policies of the government. Therefore, there is an exigency for GO to improve their functionalities in order to provide a better service to the public.

AI-driven Data Management System (DMS) can enable organizations to reduce data redundancy, prevent data losses, and prevent the recurrence of multiple input errors [11]. In the document management system, data analytics is one of the most critical functions of artificial intelligence. Adequate and comprehensive analysis of data will contribute to the growth of the enterprise. It is also essential to secure documents, susceptible ones, along with the management of records [12]. AI-powered DMS provides the highest levels of document protection by detecting confidential and personal information from documents. Data can be generated from various sources with advanced technology; however, this data needs to be analyzed and organized to gather useful information. The management and extraction of knowledge from documents will save a lot of time and effort. Documents can be clustered and analyzed by an artificial intelligence and Machine Learning (ML) framework [13]. It supports to understand the essence of a document and its relationship with other documents. Artificial intelligence technology may use optical character recognition to construct a document management system to interpret what is on a document to classify it accurately. Based on the categorization, it aids further in automating workflow processes. BD is one of the latest technologies and frameworks that can use real-time processing and analysis to maximize the value of substantial quantities of highly diverse data.

The introduction of BD has provided an opportunity for an organization to adopt multiple applications and services to enhance its performance. Recently, large quantities of data have been generated from various government outlets, such as the internet, cloud systems, and mobile devices [3][14]. Each GO is maintaining a large data corpus through BD and communicating with other organizations. In addition, employees and departments are availing of the services in the BD environment, and the size of data is increasing exponentially [3][15]. E-governance is the use of new communication and information technology to enhance organization performance, efficacy, and service delivery, and to encourage accountability [5][16][17][18]. Both BD and E-governance offers decision-makers to make a better decision and improve the organization performance that leads to customer/client satisfaction. Most organizations' mission or objective facilitates the establishment of products and services. They have to procure resources, transform them into outputs, and deliver them to their intended users.

Both AI and BD provide an opportunity for firms to understand their customers. BD offers completely new levels of exploration of hidden outcomes. In the past, organizations were unable to analyze such large amounts of data, and now that the ability to do that would lead to unintended business benefits. The consistency of the details is one more added challenge. Some documents do not even comply with the minimum clarity bar. BD defines the high amount of structured and unstructured data that floods an organization regularly. The storage of BD is distributed over several computers to reduce the cost of implementation and maintenance. Moreover, it supports the search algorithm to search in a specific memory instead of a huge memory/dataset.

Sentiment analysis utilizes Natural Language Processing (NLP) and ML in data sources to perceive and identify emotions [19]. It is also used in business to identify social data sentiment, assess the company's reputation, and understand clients. It recognizes and categorizes the opinion expressed in the text, such as news stories, content on social media, ratings, etc. Similar to many NLP techniques, it needs to be able to deal with the language complexities [20]. Organizations must understand people's emotions as clients share their thoughts and feelings more freely than ever before. The automated identification of the type of correspondence in the document centre assists a department or individual to respond in time.

Chatbots are small programs that assist users dynamically based on a set of predefined conditions, and events [19][20]. Chatbots are used for different purposes in interactive systems, including customer service, request routing, or information collection. While some chatbot implementations use comprehensive word classification processes, natural language processors, and advanced AI, others use common phrases obtained from a related library or database to search for general keywords and produce responses.

The role of KPIs within the organizational framework is to provide customizable and readily accessible measurement to users/clients to improve the quality and effectiveness of their operations [6][7]. The significance of using KPIs relies on how effectively they support better to understand the factors for an organization's success. For instance, a government company such as a university or a hospital would have different fiscal KPIs than a publicly traded corporation [7]. Each KPI should represent the organization's mission and objectives. A common mistake in creating KPIs is that a declaration such as "Improve productivity of a department" as a KPI is too general. It needs to be precise and measurable, like "Improving the individual performance of a specific department in a specific period" in order to be successful [6]. KPIs are crucial to tracking progress in publishing and reporting. Report formats should be tailored to position and purpose so that managers see a summary view, whereas department heads have much more comprehensive measurements [21]. The combination of dashboards, scorecards, informative reports, and ad-hoc research self-service tools should be considered. The KPI results allow decision-makers to determine where the emphasis should be placed to enhance the department/organization's performance [6]. In addition, Analytics serve to build a strategic plan based on the knowledge available. The creation of standardized real-time dashboards requires an investment of time and much more time to use the data effectively. Individuals will be notified of issues based on data analysis, identify potential solutions, and receive ideas for new opportunities.

In the analytical process, ML will assist, clearly identify odd trends in procedures, and send suggestions about what to do next; this process can provide useful AI insights. By developing consistent operating metrics and evaluating efficiency, businesses benefit from the use of what is necessary for being competitive on the market and provide adequate services to the customers. Key Performance Indicator (KPI) refers to a collection of quantifiable metrics used to assess the overall long-term performance of an organization. It supports to evaluate the strategic, financial, and operational achievements of a business, especially relative to those of other firms in the same industry [6][21]. Fig. 1 shows the analytical framework of the proposed research. Metadata is the primary source for this research. BD provides metadata to generate meaningful information through AI applications and KPI.

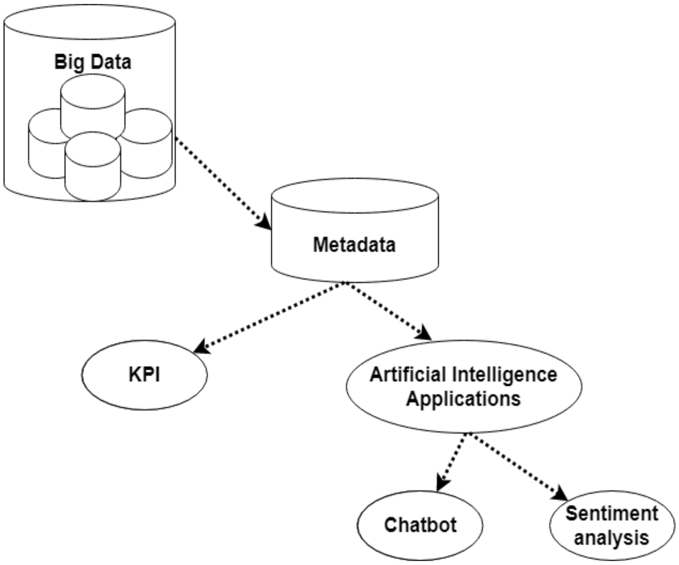


Fig. 1. Proposed Techniques for Analyzing Metadata.

The study aims to develop a framework that AI and BD concerning organization needs, which include Sentiment Analysis, Automated ChatBot, and analytical techniques for government organizations in the Kingdom of Saudi Arabia. The proposed framework support organizations to improve the quality of service rendered to the customers. It assesses organizations to achieve the objectives of vision 2030 of the Kingdom of Saudi Arabia.

Hierarchical analysis technique analyses the organization tree structure and support decision-makers/administrators to measure the performance of their department/employee. Moreover, the study discusses privacy and policy issues related to AI and BD. BD, E-Governance, and AI are the elements of a modern organization framework.

The structure of the study is organized as follows: Section 2 discusses the related works of BD, AI, and KPI in government organizations. Section 3 provides the methodology to develop KPI for BD applications. Section 4 describes the outcomes of the proposed framework, and Section 5 discusses the achievements of the proposed framework. Finally, Section 6 concludes the study with its future direction.

II. Related Works

BD analytics is stated as a method to analyze data and discover hidden trends [22]. The exponential growth of data and its analysis using different mining methods leads to a better understanding of workforce in an enterprise. E-Governance is a decision-making and implementation process that ensures a country's resources are applicable, accessible, and driven [18]. The BD framework provides an opportunity for organizations to combine heterogeneous data. Moreover, it offers individuals and corporations a global chance of combining a wide range of public databases, fraud-detection, and tax compliance activities to enhance services and apps. The effect of BD on these e-government applications supports to promote better governance [17].

In [2], the authors discussed the opportunities of implementing AI in the workplace to support administrators in making effective decisions. They presented an empirical study that evaluated and analyzed to what extent enterprises employ analytical tools for predicting business outcomes. The outcome of the study shows that an AI-based workplace supports decision-makers in complex situations.

In [3], the authors identified the challenges related to the implementation of AI-based applications to make decisions. They offered several research propositions for Information system researchers. They discussed the issues associated with the deployment of AI and replacing the manual decision-making process. In [5], a case study related to E-Governance is discussed. The authors analyzed the existing E-Governance system in Odisha, India. The findings of the case study that the deployment of data analytical tools in the workplace have improved the work environment.

AI applications such as Chatbot and Sentiment Analysis supports the customers of an organization by providing responses and required services. Organizations are effectively and efficiently managing documents through AI document analysis, data extraction, ML for content management, and Information security. BD is regarded as a reliable and useful source that can be analyzed with AI applications [14]. The research forecasts and makes decisions that may increase the reputation of the business. BD and AI are two concepts that are commonly used when describing the company's future [5][14][15]. The opportunity to implement them in various aspects of the business has captured the imagination of many, specifically, how AI can reduce the resources and provide a more secured work environment. BD and AI could adapt business processes and decisions to meet individual needs and preferences, enhancing process and decision efficiency [5][14]. Although wide advantages of BD and AI can be anticipated, several risks involve some policy considerations [11]. Governments will potentially establish general guidelines or legislation related to the use of BD and the application of AI, and the enterprise should be prepared to implement them.

In [7], the authors addressed the importance of KPI in improving the quality of services rendered to the customer. Different types of KPIs were discussed in this study. In [8], authors developed the performance indicators to operate and maintain an enterprise. The presence of a wide variety of analytics tools assists enterprises to organize their departments and employees in a better way. However, the absence of an efficient set of KPIs leads to issues such as less productivity, unorganized departmental activities, and low performance. Therefore, a demand for KPI which reveals hidden pattern via intelligent and hierarchical analysis of organizational tree structure. KPIs are essential to effective management by identifying underperforming aspects of the business as well as that demand increased resources [8]. It will report events that may signify problems that arise internally within the organization's operations or potential threats resulting from external events may provide management with rich information to improve the organization's strategies [6].

In [20], the applications of AI-based customer assistance in the online shopping portal are discussed. Developers follow two approaches to develop an AI Chatbot to identify the user's needs and respond with relevant responses. One of the development approaches is that the AI robot must respond to a user query with a restricted set of guidelines and internal structuring. A bag of words can be generated in such a way to map frequently used questions with appropriate solutions [23][24][25]. For instance, an automated banking bot asks the caller a series of questions to know what he or she needs. The AI bot will either repeat the instruction or pass the call to a human banking assistant if the caller gives an order that is out of reach. The second way to communicate with an AI chatbot is to know what a user is looking for and to deliver real-time answers based on progressive conversations or enhanced learning [2]. Although due to its complexity, this mechanism is still emerging, some apps such as Amazon's Alexa, Google Assistant, and Facebook AI-Chatbot are on the road to dynamic responses based on human behavioral and preferential traits [12][26]. Self-learning is a technique to understand the human mind and obtain enough knowledge to generate persuasive responses [27][28]. It is called Intelligent Chatbots or general AI chatbots. A general-purpose AI chatbot can understand language and continue learning based on the inputs and analyzing sentiments [29][30][31].

Based on the analysis of the related works, the study intends to develop a framework that applies AI and BD to enhance the processes in the government organizations as per the Vision 2030 of the Kingdom of Saudi Arabia. The proposed framework aims to assist end-users in deriving value from the data corpus. The implementation of an AI-based framework provides tools and techniques for decision-makers to analyze and monitor the workforce. Moreover, it secures the document by applying AI policies and report the status to the administration.

III. Materials and Methods

Synthetic metadata is developed based on the assumptions on the functionality and structure of the organizations in the Kingdom of Saudi Arabia. The following part of this section introduces the AI-based application. Table I presents the notations used in the proposed algorithms for Sentiment Analysis and KPI. The notations are used as a parameter in the following algorithm. It indicates the input required for the proposed technique to produce an output. For context, Sentiment analysis requires metadata to generate the type of correspondence. Similarly, the following notations are indicating an input or output of the proposed methods.

A. ChatBot

The primary sources for this application are correspondences, history of correspondences, and details of employees and departments. The proposed AI chatbot uses a bag of words technique to identify the users and respond to them with a relevant response. Initially, metadata related to department and employee are extracted and refined and stored as Bag of Words. ChatBot receives input from the user query and parses it to derive tokens. The query tokens are passed into the Bag of words to find the existence of relevant tokens. And, based on the frequency of the tokens, the response will be returned to the user.

B. Sentiment Analysis

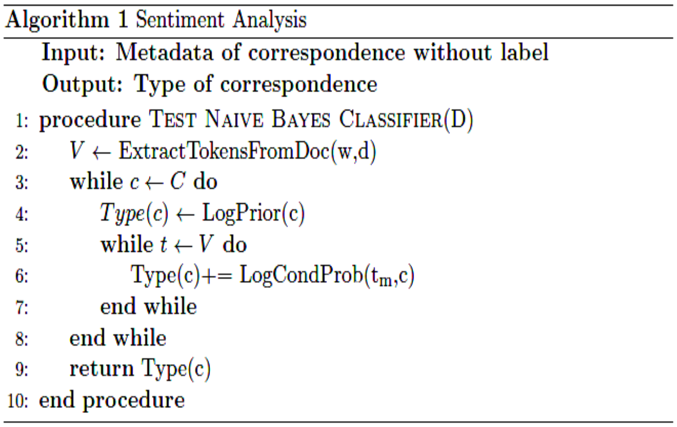
Sentiment analysis is based on the details of internal, incoming, and outgoing correspondences. The supervised learning-based Naive Bayes (NB) algorithm is employed to find a pattern from the metadata and predict the type of user correspondences. It supports users to complete correspondence in a limited time with interactive solutions. During the training phase, the labels are required to train the proposed technique. The labels are extracted from the metadata(D). It includes the type of correspondence, urgency level, and department level are used as a level. A conditional probability is generated using tokens and returned as an output with vocabulary. Fig. 2 provides the testing phase of the sentiment analysis. The testing phase able to predict the type of correspondence using the conditional probability of data. An NB classifier is used to classify the type of correspondence and helps to complete an urgent correspondence in a limited amount of time. The notations used in Fig. 2 were explained in the initial part of this section.

C. Key Performance Indicators

The organization’s tree structure is classified into four levels: Root-level, Level – 2 department, Department, and Employees. The processes in the development of KPI and classification of Department and Employees are shown in Fig. 3. And, the initial processes are representing the identification of KPI to analyze each level individually and hierarchically. The performance of the proposed KPI is measured by computing the total correspondence for Root and Level-1 departments, respectively.

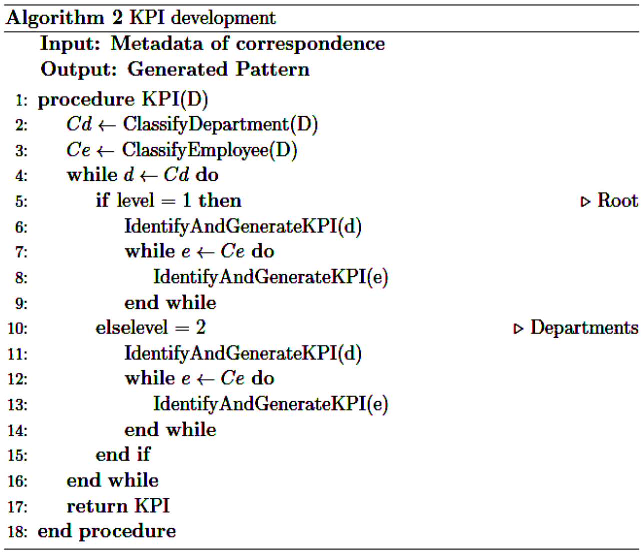
TABLE I. Notation and Meaning

Notation	Meaning	Notation	Meaning
D	Metadata	Cd	Classified departments
d	Department	Ce	Classified employees
W	Vocabulary (Initial)	e	Employees
V	Vocabulary (Testing phase)	C	Class
t_m	Term		



```
Algorithm 1 Sentiment Analysis
    Input: Metadata of correspondence without label
    Output: Type of correspondence
procedure Test Naive Bayes Classifier(D)
    $V \leftarrow$ ExtractTokensFromDoc(w,d)
    while $c \leftarrow C$ do
        $Type(c) \leftarrow$ LogPrior(c)
        while $t \leftarrow V$ do
            Type(c)+= LogCondProb($t_m$,c)
        end while
    end while
    return Type(c)
end procedure
```

Fig. 2. Proposed Algorithm for Sentiment Analysis.



```
Algorithm 2 KPI development
    Input: Metadata of correspondence
    Output: Generated Pattern
procedure KPI(D)
    $Cd \leftarrow$ ClassifyDepartment(D)
    $Ce \leftarrow$ ClassifyEmployee(D)
    while $d \leftarrow Cd$ do
        if level = 1 then                         ▷ Root
            IdentifyAndGenerateKPI(d)
            while $e \leftarrow Ce$ do
                IdentifyAndGenerateKPI(e)
            end while
        elselevel = 2                            ▷ Departments
            IdentifyAndGenerateKPI(d)
            while $e \leftarrow Ce$ do
                IdentifyAndGenerateKPI(e)
            end while
        end if
    end while
    return KPI
end procedure
```

Fig. 3. Proposed Algorithm for KPI Development.

Fig. 4 illustrates the classified departments and employees in an organization. The Root / Senior management (level – 1) for entire departments in an enterprise. Level – 2 departments are in level – 2, which indicates a cluster of departments. Departments are in level – 3 that represents single / multiple departments and finally, Employees are in level – n to mean an individual in an enterprise. Each employee may have multiple roles to represent their activities in different departments.

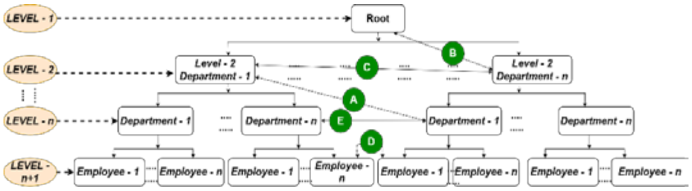


Fig. 4. Hierarchical Analysis in the Organization's Tree Structure.

Senior management has a management level as a sub-department under level -1. In the same way, each management and departments are classified with relevant sub-departments and employees. The Level – 1, senior management have access to all levels and privilege to have direct departments under it. As per the organization tree, Level – 1 is a root (Senior management), Level – 2 is the parent departments for the subsequent levels, Level – 3 is Department, and finally, Level – n is Employee.

A standard set of KPI is identified for all four levels (Root, Level-2 department, and Department, and Employee) to calculate the number of correspondences between two departments. In Fig. 4, the link "B" is used to show the transaction between Root and Level – 2 departments, i.e., level – 1 and Level – 2. For instance, suppose the Root, Level – 1 has sent a total of 17 correspondences to Level – 2 department – n of Level – 2. The information about the interactions can be extracted using the proposed framework. The link "C" is used to represent the connection between Level – 2 departments in the same Level. For context, suppose a Level – 2 department - 1 of Level – 2 has sent a total of 16 correspondences to another Level – 2 department - n of Level – 2. These details can be extracted through the proposed framework. The links "A", "D", and "E" is also used to interact with the different levels of the organization tree structure. The proposed framework provides an opportunity to extract specific correspondence between different levels as well as in the same Level.

IV. Results

In this section, the implementation of the proposed framework is discussed. Based on the functionality of the organizations, a dataset is developed similar to metadata about document transactions. The proposed KPI, Chatbot, and Sentiment analysis are implemented and tested with the synthetic dataset. The findings will not reflect the actual hierarchy of any organization.

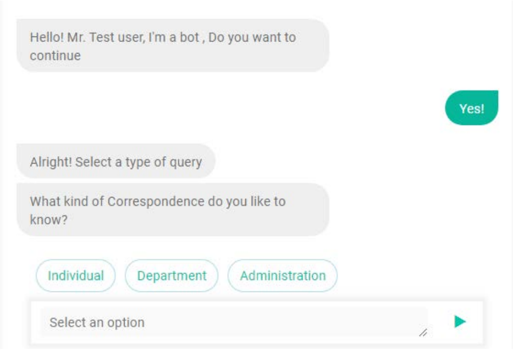
A. *Chatbot*

The researchers succeeded in the initial attempt in developing a Chatbot application using metadata that provides an automated solution to the user query. Python 3.8.1 is used for the development of Chatbot. Metadata are extracted from the database and stored in Comma Separated Value (CSV) format. The following Fig. 5 shows the initial set of results in Chatbot development. When a user logged into the application, the automated Bot has used the username and retrieved the department id and his / her last correspondence history and the total number of correspondences. The date format is represented in the Hijri date format. The proposed Chatbot uses the metadata of an organization correspondence history and correspondence forward history to retrieve the users' information. For instance, if a user entered a query "last correspondence number", then the Chatbot parses the query into tokens such as "last", "correspondence", and "number". The parsed tokens will be passed in prioritized order and fetch the responses for the user query. The further development of the Chabot application will provide a complete form of this proposed application. In this initial phase, it presents users' last correspondence details according to their username.

B. *Sentiment Analysis*

The supervised learning approach is followed in the research for developing the proposed Sentiment Analysis technique. To implement the sentiment analysis using metadata, Python 3.8.1, and Power BI are employed. The primary sources for this proposed method are correspondences (Internal, Incoming, and Outgoing) and correspondence history. Power BI is used to generate the bag of words from the subject of the correspondences for each department. Fig. 6 shows one of the bags of words generated for a department. The token that has the highest frequency is shown in elevated form rather than other tokens.

NB algorithm is applied for the identification of the type of the correspondences. The identification process supports promoting the correspondence in order to complete it in a shorter amount of duration. The development process of sentiment analysis is in progress, and the current form differentiates the sent and received correspondences with its level. For instance, if a user creates a correspondence, then the proposed system automatically predicts the type of correspondence which supports administration to complete it within a limited amount of time.



Hello! Mr. Test user, I'm a bot , Do you want to continue

Yes!

Alright! Select a type of query

What kind of Correspondence do you like to know?

Fig. 5. Proposed Chatbot Application.

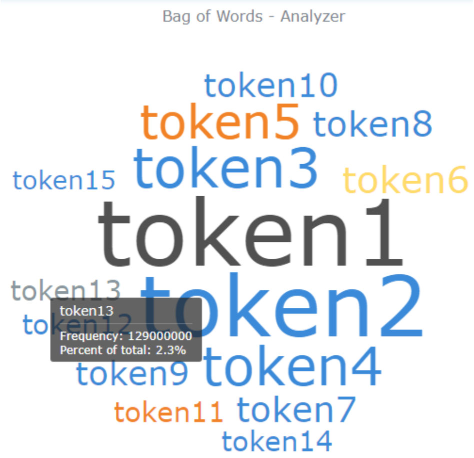


Fig. 6. Bag of Words.

C. Key Performance Indicators

A set of common KPIs for Root, Department, and Employee levels are identified via this study. The following is the list of KPIs proposed for government organizations: Root Level: Ranking workload, Estimating achievement percentage, Ranking Workload based on time, and Ranking Workload based on the number of forwards. Department Level: Estimating achievement time, Estimating achievement percentage, Actions per hour, Percentage of workload allocation, and Percentage of workload distribution. Employee Level: Estimating achievement time, Percentage of workload allocation, Estimating achievement percentage, and Actions per hour.

Oracle 11G in Windows 10 environment is employed to test the KPI and generated the results, and rendered the graphs using Canvas.Js, which provides Javascript API to render the graph. A KPI dashboard is developed to show the outcome of the KPIs as shown in Fig. 7 to 10. Fig. 7 presents the outcome of the KPI for the department level. It shows the performance of a specific department in an organization.

Fig. 7 and 8 illustrates the graph rendered via the proposed KPI for the department level. It supports the decision-maker to know that the achievement percentage and time of a department. Thus, the decision-maker will decide to enhance the performance of both departments and employees to complete the correspondences in time to improve the achievement percentage and time.

Fig. 9 shows the KPI insight for the employee level. For context, an employee received a sum of 1416 correspondences in the year 1440. The performance of that employee, action per hour is shown on the right side of Fig. 9. Fig. 10 shows the percentage of achievement for the employee. The employee has taken an average of 1.33 days to complete a correspondence that may affect the organization's overall performance.

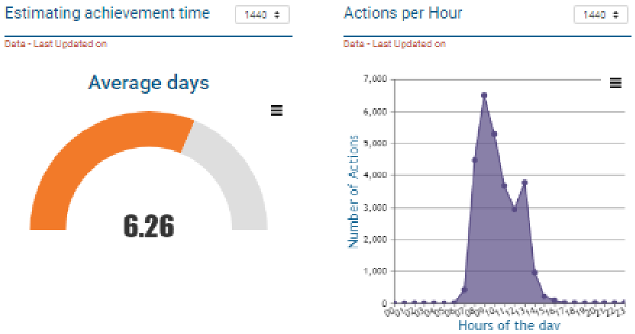


Fig. 7. KPI – Department Level – 1.

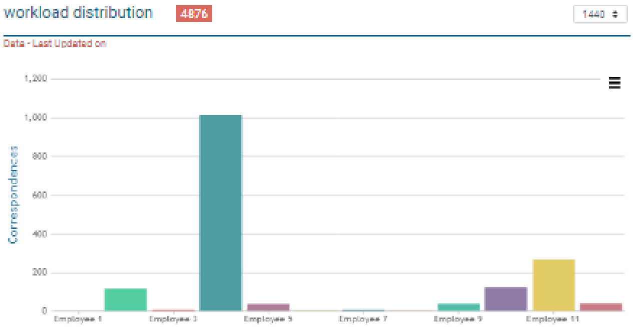


Fig. 8. KPI – Department Level – 2.

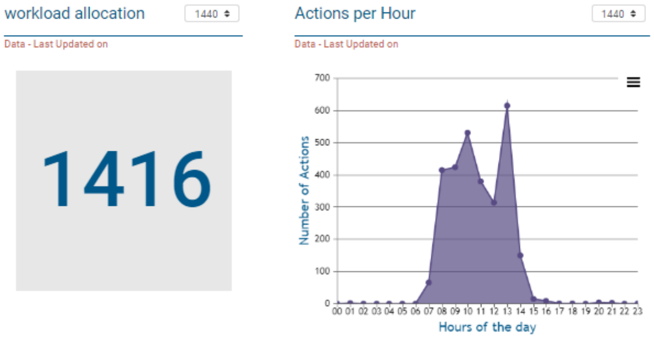


Fig. 9. Employee Level – 1.

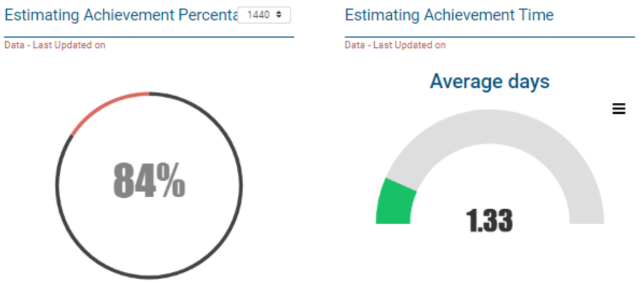


Fig. 10. Employee Level – 2.

V. Discussions

The outcome of this study shows that metadata of a correspondence/document produces a sufficient amount of information for decision-makers to make an effective decision. More and more data linked to documents, the more alternatives are open for extracting observations, facts, and figures that assist in monitoring and tracking the workforce and complete a task in time. Such kind of data is a strong foundation and an essential element for the launch and implementation of high-quality indicators. The National Centre for Archives and Records in Saudi Arabia has published a range of laws and policies on all kinds of documents that would strengthen administrative work and transactions. It includes an adequate database to store data concerning statistics and operational management indicators.

The results of this research show that the proposed framework can present an effective environment for an organization to manage its workforce. Both Chatbot and Sentiment Analysis support the organizations to respond to their clients' query automatically. Moreover, the KPI dashboard assists decision-makers in planning without any difficulties.

Operational and administrative processes share several characteristics [23]. Both involve sequences of linked, interdependent activities that together transform inputs into outputs. Both have clients who may be internal or external to the organization—the primary differences between the two lies in their outputs [24]. Typically, operational processes produce goods and services that external customers consume, while administrative procedures generate information and action plans for the internal groups [8]. For this reason, the two are frequently considered independent, unrelated activities, even though they must usually be aligned and mutually supportive if the organization is to function effectively.

The viewpoint of the work processes has contributed to a range of significant perspectives for executives [24]. It offers a beneficial context for a common organizational issue to be addressed: fragmentation or lack of integration. Many aspects of modern organizations, including complexity, highly segregated sub-units, and responsibilities, weak informal relationships, scale, and physical distance, make integration difficult [24]. The simple identification of work processes as viable units of study and goals of managerial action also enhances integration. For illustration, charting hierarchical workflow or pursuing an order through the fulfilment method are convenient ways to inform employees that the operations of various departments and regional units are interdependent, even though organizational charts indicate that beyond their vertical lines of authority [10].

Moreover, the viewpoint of the work processes offers fresh objectives for change. Managers discuss the fundamental mechanisms rather than concentrating on systems and functions. An apparent benefit is that they can analyze the organization's actual work closely.

Most of the improvement programs concentrated primarily on the restructuring of systems; the continued implementation and maintenance of the reconfigured systems were largely ignored [9][25]. However, without enough monitoring, coordination, and control, as well as periodic interference, even the best processes will not perform effectively. Furthermore, organizational processes were typically aimed at enhancement, although their enabling administrative functions were ignored [8]. Incompatibilities and contradictions occurred because there was no knowledge and preparations required for an efficient operation. The implementation of the Chatbot and Sentiment Analysis technique will support an organization to minimize the resources and satisfy the user requirements. The combination of hierarchical and traditional KPI can offer a better outcome to the organization to improve their performance.

The principles of information security support organizations to ensure unauthorized access to records [28]. For the proposed framework, some security concepts are framed as follows:

- Documents produced and accepted using this standard should be inspected for modifications, updates, or retirement of documents at regular intervals [29]. During regular document maintenance, the existing document owners must be verified or affirmed. However, it does not exceed three years from the previous revision date when the document is preserved ordinarily [30].
- The document developer must identify the process owners, service owners, and stakeholders and analyze the document during the development / initial analysis [30]. Although there is no prerequisite for any person or community review, it is highly recommended that members of those most affected by the content of the document should take part in the study.
- DMS must allow administrators to establish system permissions that are regulated based on the individual level of the document, ensuring that only approved staff has access to system files and documents [31]. Thus, confidential data is not in the hands of the wrong person, and the possibility of data breaches is greatly reduced.

As per the outcome of this study, the proposed Chatbot supports users to understand the status of their correspondence. It uses an AI approach to read the user query and provide a necessary response. The focus of the Chatbot is to minimize manual interactions with users. It works similar to an automated Chatbot in a banking system that responds to user queries. In the proposed Chatbot, the user needs to select an option to retrieve a response from BD. The proposed method uses a customized dataset depends on the employee and department. For context, an employee of level 2 requires his/ her correspondence information; the Chatbot searches his / her query in the level 2 dataset, not in the entire dataset. In case, data is not found in level 2; then the query will be matched with its higher level, i.e. level 1.

The proposed Sentiment analysis technique is able to predict the type of correspondence and support the decision-maker / administrator to complete the transaction in a limited amount of time. In the existing model, the user needs to provide the type of correspondence. There is a possibility of error in this model. For instance, for an urgent task, the user may select a standard type instead of a critical type that leads the organization to postpone the urgent task to a later time. The proposed framework is capable of predicting the nature of correspondence automatically and assist the organization in increasing the achievement rate and turnaround time. The further development of this framework will lead to finding solutions for more challenges in a complex work environment.

The proposed research recommends implementing benchmark data derived according to the flow of documents to recognize best practices and growth opportunities. It is necessary to hold frequent conferences to enhance and exchange information related to archiving documents that will improve the role of IT in the archive service to become an intelligent centre and support decision-making processes. Moreover, developing a universal standard for archiving center supports administrative experts to construct, analyze, and plan for successful indicators. Before archiving a document and related data, a review must be performed to ensure the data supports the objectives and vision of the organization. Developing personalized and specialized courses can lead to a growth of professional human resources dealing with electronic documents and transactions.

Policymakers have an interest in AI. The three broader areas in which policymakers seek to ensure that organizations strike an adequate balance between their business interests and customers' interests includes: data protection, prejudice, and ethics [11][13][27]. Decision-makers need insights into how the data are collected to better understand and resolve privacy issues at present and how to handle failures in data privacy, especially with rules and regulations of National Archive and Records Centre, Kingdom of Saudi Arabia and Information security principles. Many AI algorithms are opaque black boxes, exacerbating the problem, and therefore it is hard to isolate the exact factors considered by these algorithms [25]. AI Organizations and developers must contend with ethics. Privacy decisions may represent the policy of an organization and could also be motivated by ethical concerns. In this context, research should look at "how regulatory ethics can pave the way for organizations to exceed consumer privacy expectations and oversubscribe to legal mandates to preserve self-regulation" [4]. A similar focus of research may be to explore how ethical questions regarding AI differ across cultures.

Finally, the proposed research indicates employees in the document centre to apply restrictions for protecting data during transactions. Data protection is one of the main steps towards ensuring the sustainability of the data. The protection processes should be initiated along with the creation of a correspondence. Data entered by users should be reviewed in order to maintain the correct format. It is also necessary to review the correspondence data to ensure that it supports the vision and mission of the organization to reach an optimal solution in the processes of data protection.

VI. Conclusion

The kingdom's vision of 2030 has initiated many development programs in several areas and measure the same. In the recent Global Summit AI, 2020, the new strategy for Data and AI is announced for implementing AI in the government and private sectors. The emergence of BD and AI techniques support organizations to identify the pitfalls that affect the performance and find a solution for improvement. The documents and their flows provide the documents and archives centre with an opportunity to make effective decision-making. Data granularity can feed information into a range of future activities and events. Since an organization is focused on predicting how risk is realized, having access to BD can change the entire working model of an enterprise/organization. However, the granularity of data may also contribute to the promotion of risk classification, where achievement rates are focused on a group of individuals with similar profiles. One of the statistical methods for representing information about an individual entity of an organization is the key performance indicator. An employee's success will significantly increase the performance of the relevant department. Therefore, it is necessary to evaluate individual performance and how it is beneficial to enhance an organization's overall performance. In this study, an AI-based Chatbot and Sentiment analysis were developed to support users to follow-up their correspondences. It has identified some useful key performance indicators that can measure the performance of employees at different levels. The initial results from Chatbot and Sentiment analysis are promising, and further development will provide a complete dimension to these applications. Also, the study has formulated and discussed the information security and AI policies to develop and maintain a document in the document centre.

References

[1] Saudi Vision 2030, “www.vision2030.gov.sa,” 2017.

[2] E. Nica, Miklencicova Renata, and Kicova Eva, “Artificial intelligence-supported workplace decisions: big data algorithmic analytics, sensory and tracking technologies, and metabolism monitors,” Psychosociological Issues Human. Resource Management, vol. 7, no. 2, pp. 31–36, 2019.

[3] Y. Duan, J. S. Edwards, and Y. K. Dwivedi, “Artificial intelligence for decision making in the era of big data – evolution, challenges and research agenda,” International Journal of Information Management., vol. 48, pp. 63–71, Oct. 2019, doi: 10.1016/j.ijinfomgt.2019.01.021.

[4] K. D. Martin and P. E. Murphy, “The role of data privacy in marketing,” Springer, no. 45, pp. 135–155, 2017, doi: 10.1007/s11747-016-0495-4.

[5] P. Patnaik, S. Pattnaik, and P. Singh, “Use of data analytics for effective E-Governance: a case study of ‘EMutation’ system of odisha,” in Lecture Notes on Data Engineering and Communications Technologies, vol. 37, Springer, 2020, pp. 109–122.

[6] D. A. Bishop, “Key performance indicators: ideation to creation,” IEEE Engineering Management, Rev., vol. 46, no. 1, pp. 13–15, Mar. 2018, doi: 10.1109/EMR.2018.2810104.

[7] M. Badawy, A. A. A. El-Aziz, A. M. Idress, H. Hefny, and S. Hossam, “A survey on exploring key performance indicators,” Future Computing Informatics Journal., vol. 1, no. 1–2, pp. 47–52, Dec. 2016, doi: 10.1016/j.fcij.2016.04.001.

[8] J. H. K. Lai and C. S. Man, “Performance indicators for facilities operation and maintenance (Part 2): Shortlisting through a focus group study,” Facilities, vol. 36, no. 9–10, pp. 495–509, Jul. 2018, doi: 10.1108/F-08-2017-0076.

[9] T. Dewett and G. R. Jones, “The role of information technology in the organization: a review, model, and assessment,” Journal Management, vol. 27, no. 3, pp. 313–346, Jun. 2001, doi: 10.1177/014920630102700306.

[10] N. K. Dev, R. Shankar, R. Gupta, and J. Dong, “Multi-criteria evaluation of real-time key performance indicators of supply chain with consideration of big data architecture,” Computing Industrial

Engineering, vol. 128, pp. 1076–1087, Feb. 2019, doi: 10.1016/j.cie.2018.04.012.

[11] T. Davenport, A. Guha, D. Grewal, and T. Bressgott, “How artificial intelligence will change the future of marketing,” Journal Academy Marketing Sciences, vol. 48, no. 1, pp. 24–42, Jan. 2020, doi: 10.1007/s11747-019-00696-0.

[12] D. Valle-Cruz, R. Sandoval-Almazan, E. A. Ruvalcaba-Gomez, and J. Ignacio Criado, “A review of artificial intelligence in government and its potential from a public policy perspective,” in ACM International Conference Proceeding Series, Jun. 2019, pp. 91–99, doi: 10.1145/3325112.3325242.

[13] T. Davenport, “Enterprise analytics: Optimize performance, process, and decisions through big data,” 2013.

[14] C. Moreno, R. A. Carrasco, and E. Herrera-Viedma, “Data and Artificial Intelligence Strategy: A Conceptual Enterprise Big Data Cloud Architecture to Enable Market-Oriented Organisations,” International Journal of Interactive Multimedia and Artificial Intelligence, vol. 5, no. 6, p. 7, Jun. 2019, doi: 10.9781/ijimai.2019.06.003.

[15] D. A. McFarland and H. R. McFarland, “Big data and the danger of being precisely inaccurate,” Big Data Society, vol. 2, no. 2, 2015, doi: 10.1177/2053951715602495.

[16] M. Kowalczyk, “Key performance indicators in local government in Poland,” Pr. Nauk. Uniw. Ekon. we Wrocławiu, no. 503, pp. 236–245, 2018, doi: 10.15611/pn.2018.503.21.

[17] A. Hooda and M. L. Singla, “Reengineering as a strategic stance for e-governance success - mediating role of core competencies: a mixed method study,” Transform. Gov. People, Process Policy, vol. 14, no. 2, pp. 205–235, May 2020, doi: 10.1108/TG-01-2020-0017.

[18] M. Y. Febrianta and H. Amani, “Identification of e-Governance indicators for measuring smart governance in bandung city,” 2019, doi: 10.31227/osf.io/avbsu.

[19] X. Luo, S. Tong, Z. Fang, and Z. Qu, “Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases,” Mark. Sci., vol. 38, no. 6, pp. 937–947, Sep. 2019, doi: 10.1287/mksc.2019.1192.

[20] E. Pantano and G. Pizzi, “Forecasting artificial intelligence on online customer assistance: Evidence from chatbot patents analysis,” Journal of Retail Consumers Serv., vol. 55, p. 102096, Jul. 2020, doi: 10.1016/j.jretconser.2020.102096.

[21] D. Agostino and M. Arnaboldi, “Rational and ritualistic use of key performance indicators in hybrid organizations,” Public Money Management, vol. 37, no. 6, pp. 409–416, Sep. 2017, doi: 10.1080/09540962.2017.1344021.

[22] P. C. Verhoef et al., “Consumer Connectivity in a Complex, Technology-enabled, and Mobile-oriented World with Smart Products,” Journal of Interactive Marketing, vol. 40, pp. 1–8, Nov. 2017, doi: 10.1016/j.intmar.2017.06.001.

[23] N. Gonzalez et al., “A quantitative analysis of current security concerns and solutions for cloud computing,” Journal of Cloud Computing, vol. 1, no. 1, pp. 1–18, Jul. 2012, doi: 10.1186/2192-113X-1-11.

[24] E. Girdzijauskaitė, A. Radzevičienė, and A. Jakubavičius, “Impact of international branch campus KPIs on the university competitiveness: FARE method,” Insights into Reg. Dev., vol. 1, no. 2, pp. 171–180, Jun. 2019, doi: 10.9770/ird.2019.1.2(7).

[25] S. Mukherjee, “How IT allows E-participation in policy-making process,” arXiv:1903.00831, Mar. 2019, doi: 10.6084/m9.figshare.7796063.v2.

[26] A. Agrawal, J. Gans, and A. Goldfarb, “Economic policy for artificial intelligence,” Innov. Policy Econ., vol. 19, no. 1, pp. 139–159, Oct. 2019, doi: 10.1086/699935.

[27] J. Jung, R. Shroff, A. Feller, and S. Goel, “Algorithmic decision making in the presence of unmeasured confounding,” arXiv Prepr. arXiv1805.01868, May 2018, Accessed: Sep. 21, 2020. [Online]. Available: http://arxiv.org/abs/1805.01868.

[28] S. Choi, J. T. Martins, and I. Bernik, “Information security: listening to the perspective of organisational insiders,” J. Inf. Sci., vol. 44, no. 6, pp. 752–767, Dec. 2018, doi: 10.1177/0165551517748288.

[29] J. Roberts, “Organizational ignorance: towards a managerial perspective on the unknown,” Management Learning, vol. 44, no. 3, pp. 215–236, Jul. 2013, doi: 10.1177/1350507612443208.

[30] Z. Hussain, A. Taylor, and D. Flynn, “A case study of the process of achieving legitimation in information systems development,” Journal Inf. Sci., vol. 30, no. 5, pp. 408–417, 2004, doi: 10.1177/0165551504046725.

[31] L. Rasmussen and H. Hall, “The adoption process in management innovation: a knowledge management case study,” J. Inf. Sci., vol. 42, no. 3, pp. 356–368, 2015, doi: 10.1177/0165551515625032.

iDietScore^{TM}: Meal Recommender System for Athletes and Active Individuals

Norashikin Mustafa[1]
[1]Faculty of Health Sciences
Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia
[1]Department of Nutrition, Kulliyyah of Allied Health Sciences, International Islamic University Malaysia
Kuantan, Malaysia

Abdul Hadi Abd Rahman[2]*, Nor Samsiah Sani[3]
Center for Artificial Intelligence Technology
Universiti Kebangsaan Malaysia, Bangi, Malaysia

Mohd Izham Mohamad[4], Ahmad Zawawi Zakaria[5]
National Sport Institute
Kuala Lumpur, Malaysia

Azimah Ahmad[6]
National Defense University of Malaysia
Kuala Lumpur Malaysia

Noor Hafizah Yatiman[7]
Ruzita Abd Talib[8], Poh Bee Koon[9]
Nutritional Science Program and Centre for Community Health, Faculty of Health Sciences, Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia

Nik Shanita Safii[10]
Dietetics Program and Centre for Community Health
Faculty of Health Sciences, Universiti Kebangsaan Malaysia
Kuala Lumpur, Malaysia

Abstract—**Individualized meal planning is a nutrition counseling strategy that focuses on improving food behavior changes. In the sports setting, the number of experts who are sports dietitians or nutritionists (SD/SN) is small in number, and yet the demand for creating meal planning for a vast number of athletes often cannot be met. Although some food recommender system had been proposed to provide healthy menu planning for the general population, no similar solution focused on the athlete's needs. In this study, the iDietScore^{TM} architecture was proposed to give athletes and active individuals virtual individualized meal planning based on their profile, includes energy and macronutrients requirement, sports category, age group, training cycles, training time and individual food preferences. Knowledge acquisition on the expert domain (the SN) was conducted prior to the system design through a semi-structured interview to understand meal planning activities' workflow. The architecture comprises: (1) iDietScore^{TM} web for SN/SD, (2) mobile application for athletes and active individuals and (3) expert system. SN/SD used the iDietScore^{TM} web to develop a meal plan and initiate the compilation meal plan database for further use in the expert system. The user used iDietScore^{TM} mobile app to receive the virtual individualized meal plan. An inference-based expert system was applied in the current study to generate the meal plan recommendation and meal reconstruction for the user. Further research is necessary to evaluate the prototype's usability by the target user (athletes and active individuals).**

Keywords—Expert system; meal planning; sports nutrition; inference engine; design and development

I. INTRODUCTION

Athletes need adequate energy and nutrition as fuel to sustain their long training hours and maintain their health [1]. Understanding an athlete's training periodization plan would give an idea or guideline for dietitians or nutritionists to match the nutrition strategies to support the training outcome [2]. Athlete training is divided into different cycles throughout the years and each of the cycles consists of different volume, frequencies and intensity of training sessions. Therefore, food for athletes should also change to meet different nutrition demands [3]. Several cross-sectional studies on athlete's dietary intake found that most of them did not meet their energy requirements during training and competition [4]–[7]. Besides, a systematic review identifies that most of the semi-professional and professional team sports athletes exceed the needs of protein and fat during training and competition [6]. Inadequate nutrition intake not only occurred among adult or elite athletes but also affected young athletes. A systematic review by reference [8] identified that adolescent athletes (age 10-19 years old) did not adjust their nutrient intake based on their sport and intensity of training. Low energy intake among athletes may lead to several health consequences such as loss of muscle mass; menstrual dysfunction; loss of or failure to gain bone density; an increased risk of fatigue, injury, and illness; and a prolonged recovery process [1]. This condition may affect an athlete's carrier, performance and health. Therefore, action needs to be taken to improve athletes' dietary intake, especially during training and competition.

Meal planning is one of the nutrition counseling strategies that facilitate food behavior changes. Meal planning is a detailed meal plan listing precisely the type of food with the portion size to be eaten [9]. Moreover, meal planning is viewed as one technique to deliver nutrition knowledge in a more practical way [10]. According to four randomized control trial studies, preparing the meal plan was a helpful strategy in achieving health and food behavior changes among middle-aged adults [9]. An expert's knowledge of food composition, usually by a dietitian or a nutritionist, is needed to translate nutrition prescription into food choice and mealtime [11]. In the sports setting, the number of experts (sports dietitians or nutritionists) are small, and the demand for creating meal planning for a huge number of athletes often cannot be met. Moreover, traditional meal planning development using pen and paper is time-consuming.

Considering the fact that advanced technology may be used to assist people in improving health, the current study proposed an architecture design of iDietScoreTM, a system that provides virtual meal plans based on athlete's or active individual's profiles (include food preference) and expert's suggestion. The expert system provides a good platform for implementing technologies that may be identical or comparable to human experts. In this study, sports nutritionists who had domain knowledge of food options to produce the equivalent macronutrient meal planning for their athletes. Thus, athletes and active individuals able to receive meal planning at any time and location, especially when sports dietitians or nutritionist is not available. The present paper is organized as follows: Section 2 presents the related work; in Sections 3 and 4 the system design and development are described and finally, conclusions and future work are drawn in Section 5.

II. Related Works

Expert systems provide an excellent platform to implement applications that can be similar or near to human experts, such as diagnosing and assisting humans in decision-making, suggesting an alternative option to a problem and advising [12]. In recent years, the expert system's nutrition and balanced food domain has been discovered as a possible solution to direct the user to meet their personal nutrient needs [13]–[16]. Meal recommender or meal planning is considered as a multi-dimensional problem since it includes several decision variables with multiple constraints and objectives. In general, most of the study aims to develop a meal recommender based on nutrition recommendation [17] and recent studies include 'user's food preference [13][18][19]. Food preference is the key element of personalizing nutrition. Personalized nutrition has been defined in a number of ways, and this research describes it as an environment that empowers human autonomy to drive nutrition strategies that prevent, manage and treat diseases and improve health [20]. According to characteristics described by reference [21], personalized nutrition not only act as a disease preventive tools but it also empowers individuals to make a healthy choice according to their preferred foods and characteristics. Determine individual food preference is quite challenging as it depends on many factors such as culture, religion, knowledge and food availability [13], [22], [23]. A less palatable combination of food or unfamiliar food that suggests to the user might lead to non-adherence of the recommendation. Thus, the local food database is a crucial component to be included in the meal recommendation system.

In general, most studies with customized diet recommendations have few layers to process the information before the final recommendation. These layers include information gathering, user profile dataset, the intelligent system and the end-user interface [24]. The intelligence system usually focuses on receiving input from a user profile and produce output information of the recommended meal plan. The inference engine is one of the artificial intelligence techniques used in an expert system that applies a rule-based reasoning approach into the knowledge base in order to deduced recommendations [13], [25], [26]. The domain knowledge and rules were used to generate a recommendation. The benefit of rule-based reasoning is that it can solve the data shortage or cold start issue with machine learning and collaborative filtering approach. In addition, another advantage of the rule-based system is it has uniformity of knowledge format [13].

Knowledge acquisition is an essential process of the expert system, and it is quite challenging and time-consuming, but massive of information can be collected if the appropriate method is applied [27]. Knowledge can be acquired from different sources such as experts, book and documents [28][25]. The previous study had conducted knowledge acquisition in various techniques such as interview the domain expert, review the literature, document, guideline or related web site and observation [17], [25], [27], [29], [30]. A combination of interview and observation is recommended for acquired tacit and explicit knowledge [28]. Less research on the meal recommender for athletes is discussed. Reference [31] develops a suitable system for active individuals by providing a workout session and diet plans. However, nutrition rules for athletes did not include in this study. Moreover, reference [3] describes the development of a personalized food and nutrition ontology working with a rule-based knowledge framework to provide specific menus for the 'weightlifter's diary nutritional needs and personal preferences. However, this system was developed only for a single type of sport.

Therefore, there is still huge potential and opportunity to explore more on developing a system that specifically for athletes or active individuals. Sports dietitians or nutritionists would be the most suitable experts for knowledge acquisition purposed in the sports nutrition domain. Thus, the proposed approach in this study belongs to the rule-based approaches. Therefore, a rule-based approached was used in the current expert system to represent human expert knowledge.

III. System Design

A. Knowledge Acquisition

Knowledge acquisition was conducted to understand the process and workflow on how sports dietitians/ nutritionists (SD/SN) translate the athlete's information and profile into individualized meal plan. Knowledge acquisition was conducted among SD/SN currently working with national athletes in Malaysia through face to face interviews. The duration of an interview session was approximately 30 to 45 minutes. A semi-structured interview was conducted to give the participants room to answer the questions. Moreover, probes were used to explore the answers provided in-depth. The semi-structured interview guide were asked about the conditions that required meal planning and to describe the processes involved during the development of a meal plan. Probe questions such as "Can you explain further?" and "Following that, what else did you do?" were asked. The interviews were audio-recorded and transcribed verbatim. The transcripts that had been produced were then shared with the participants to check for the description's accuracy and adequacy. The validation of transcripts was important to make sure that the researcher's account truly reflected the true conversation [32] and to manage the issue of reliability or trustworthiness [33]. A thematic analysis was conducted and Atlas.ti 8 was used to support the labeling and retrieval of data that had been assigned a particular code [34]. This study adopted Braun and Clarke's (2006) step-by-step guidelines to create meaningful themes [35].

Table I presents six themes that emerged based on the interview and these themes were the general process that is involved in the development of meal planning for athletes practiced by SN. The sub-themes were the specific components that are important to be included in each theme or process in meal planning development.

B. *Architecture of iDietScore™*

Based on the acquisition of expertise, high critical thinking and evidence-based practice relating to sports nutrition were required during the process of planning an athlete's meal plan. Meal planning is designed to include food options consistent with athletes' nutritional needs, training schedule and dietary preferences, as illustrated in Fig. 1. Expert systems provide a good platform for implementing technologies that may be identical or comparable to human experts, such as diagnosing, assisting people in decision-making, recommending solutions to a problem and offering advice [12]. The current expert system aims to provide a virtual meal plan based on nutrition needs, training plan, training time, and food preferences for athletes and active individuals. In order to achieve the aim, an architecture of iDietScore™ system (Fig. 2) was designed based on the workflow practiced by the SN in developing individualized meal planning. The interrelated structure of iDietScore™ comprises: (1) iDietScore™ web for sports dietitians or nutritionists, (2) mobile application for athletes and active individuals and (3) expert system.

The flow starts from the collection of meal plan database from SN using the iDietScore™ web. Next, using the iDietScore™ mobile app, users must provide input on the profile page such as measurement of anthropometries, sports type, training cycle, training time, food preferences and food allergies. Based on the information, the system generates energy and nutrition requirement for the user. The expert system (ES), consisting of an inference engine (component 1), matches the user profile with a meal plan database by followed the meal plan rules that had been embedded in a knowledge base (component 2). ES proposes a meal plan that matches the user profile. Besides, ES also allows users to make changes in each food item in the meal plan by following meal reconstruction rules embedded in the knowledge base. All changes were recorded and save as a new meal plan. The descriptions of each architectural structure are addressed in the next sections that start with the iDietScoreTM web for sports dietitians or nutritionists and followed by mobile application for athletes and active individuals and expert systems.

TABLE I. SUMMARY OF THEMES AND SUBTHEMES

Themes (General Process)	**Sub-themes (specific process)**
Collecting pertinent data	- Conducting body composition assessment - Identify training periodization plan - Identify training time - Identify food and nutrition-related history
Analyzing the collected data	- Analyzing body composition - Analyzing dietary intake
Determining nutrition prescription	- Calculating energy requirement - Determining macronutrient distribution based on g/kg body weight - Using food exchange distribution table to distribute macronutrient across the mealtimes
Formulating goals and determining actions	- Determining the use of supplements - Emphasizing in gradual dietary changes strategy - Setting achievable goals - Conducting one to one meeting between SNs and athletes to discuss the meal plan
Recommending and implementing action	- Dietary education - Adjusting and improvising current dietary intake Determining mealtimes (main meal, pre & post-exercise meal) to match with training time
Monitoring	- Monitoring dietary intake - Monitoring body composition

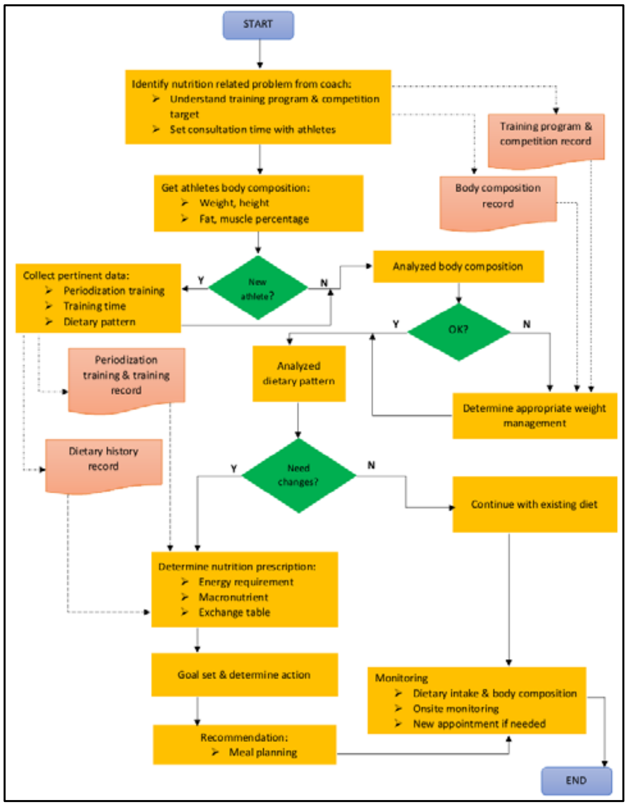

START
Identify nutrition related problem from coach:
Understand training program & competition target
Set consultation time with athletes
Training program & competition record
Get athletes body composition:
Weight, height
Fat, muscle percentage
Body composition record
Collect pertinent data:
Periodization training
Training time
Dietary pattern
Y
New athlete?
N
Analyzed body composition
Y
OK?
N
Periodization training & training record
Analyzed dietary pattern
Determine appropriate weight management
Dietary history record
Y
Need changes?
N
Continue with existing diet
Determine nutrition prescription:
Energy requirement
Macronutrient
Exchange table
Goal set & determine action
Monitoring
Dietary intake & body composition
Onsite monitoring
New appointment if needed
Recommendation:
Meal planning
END

Fig. 1. The Workflow of Meal Planning Activities as based on Interviewing Sports Nutritionist in National Sports Institute, Malaysia.

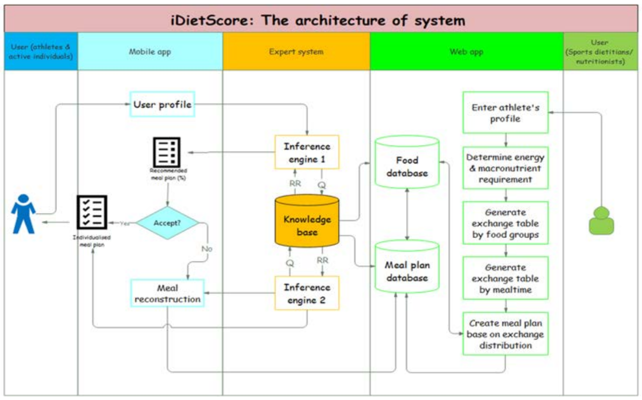

iDietScore: The architecture of system
User (athletes & active individuals)
Mobile app
Expert system
Web app
User (Sports dietitians/ nutritionists)
User profile
Recommended meal plan (%)
Accept?
Yes
No
Individualised meal plan
Meal reconstruction
Inference engine 1
RR
Q
Knowledge base
Q
RR
Inference engine 2
Food database
Meal plan database
Enter athlete's profile
Determine energy & macronutrient requirement
Generate exchange table by food groups
Generate exchange table by mealtime
Create meal plan base on exchange distribution

Fig. 2. A Design of iDietScore^{TM}.

C. *iDietScore^{TM} Web App for Sports Dietitians or Nutritionist (SD/SN)*

The traditional meal planning method with pen and paper takes time and lacks documentation. Thus, comprehensive meal planning sets cannot be compiled and reused for similar cases. iDietScore^{TM} web has been developed to assist SD / SN plan a complete set of 1-day meals for athletes and active individuals that can be compiled into meal plan database. Moreover, the web also aims to initiate the compilation of meal plan database using a web application. The meal plan database is one of the important components in the development of the expert system for iDietScore^{TM}. It was design based on current practices by SNs (knowledge acquisition, Fig. 1), sports nutrition guidelines [36] [1] and food exchange list with macronutrient content by [37].

The web automatically calculates calories (in kcal) and macronutrient (in percentage, gram/day and food exchange distribution) requirements based on sports categories (such as endurance, intermittent/power strength, skill and active individual). The energy requirement was determined based on a formula calculation of two parameters which are basal metabolic rate (BMR) and physical activity level (PAL) (Energy requirement = BMR x PAL) [38]. Thus, to come out with the requirement, input from SD/SN is still needed. SD needs to enter the user profile, verify the calculated calories recommendation, suggest suitable carbohydrate and protein intake, distribute the calculated food group exchange into mealtime and suggest appropriate food from the food database. The input from SD/SN to develop meal planning were illustrated in Fig. 3. The meal plan is saved as a whole set that is linked to the profile, such as total calories, sports categories, training time, season, food allergies and food preferences.

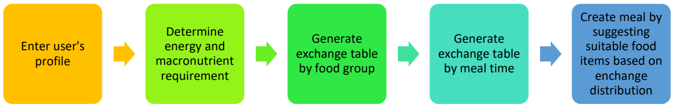

Enter user's profile
Determine energy and macronutrient requirement
Generate exchange table by food group
Generate exchange table by meal time
Create meal by suggesting suitable food items based on enchange distribution

Fig. 3. A The required Input from SD/SN to Develop Meal Plan.

D. *iDietScore^{TM} Mobile App Features and Rule-Base Expert System*

The aim of iDietScore^{TM} mobile app development is to assist the user who are athletes and active individuals to meet their calorie and nutrient requirements by suggesting them with the individual meal plan. The individual meal plan is based on their current nutrient needs, sports type, training time, training cycle, food allergies and food preferences. In addition, the user can also change the food that is suggested in the meal plan but within the control nutrient values. The rule-based expert system (ES) was type of ES that being develop in current study to generate the recommendation of meal plan and meal reconstruction for the user. The rule-based ES comprises of three main components which are user interface, inference engine and knowledge base. All information about user profiling (such as age, gender, weight, height, type of sports, training time, food allergies and food preferences) was collected at the user interface. Those information were essential for energy calculation and macronutrient (carbohydrate and protein) recommendation. The calculation was similar as described in iDietScore^{TM} web.

The next component is the knowledge base that contains the specialized knowledge of the domain problem. The current study includes rules related to individualized meal plan together with rules for meal reconstruction. All the rules were acquired from the experts (sports nutritionists), sports nutrition position statement and nutrition guideline represented in the declarative form of "if…. then…" rule. This study implements forward chaining as the inference engine follow the chain of conditions or rules to deduce the outcome. This study has two inference engines to differentiate between expert inference engine for meal plan (E1) and inference engine for meal reconstruction (E2). The 1st rule involves meal planning recommendations. Referring to the architecture (Fig. 2), upon receiving the profile input from the user, inference engine 1 (E1) will infer with meal plan recommendation (at least one meal plan) together with the score of accuracy. The accuracy of the meal plan suggested by the ES is seen in the percentage. The more rules that are followed, the greater the quality of that meal plan. The indicator will offer users a view of how reliable the meal plans to meet their nutrient needs. There are five rules for meal planning suggestions that are included in the knowledge-based. The flow chart in Fig. 4 shows how the rules (label with the alphabetic start from A until E) are being applied in the inference engine 1 (E1) to produce a meal plan suggestion to a specific user.

A higher score (50%) will be given as the meal plan meets the energy requirement (Rule A). Next, a score of 30% was given as the meal plan meets the sports categories' rule. Sports categories resemble the macronutrients distribution thus, meeting this rule will be receiving a more accurate meal plan. A score of 10% was given as the meal plan meet the rule that related to training time. Training time is related to mealtime distribution. Next, another 5% was given as the meal plan meet each of the rules related to food allergies and food preferences. The total of 100% would refer to the most accurate meal plan or meet all rules for meal plan.

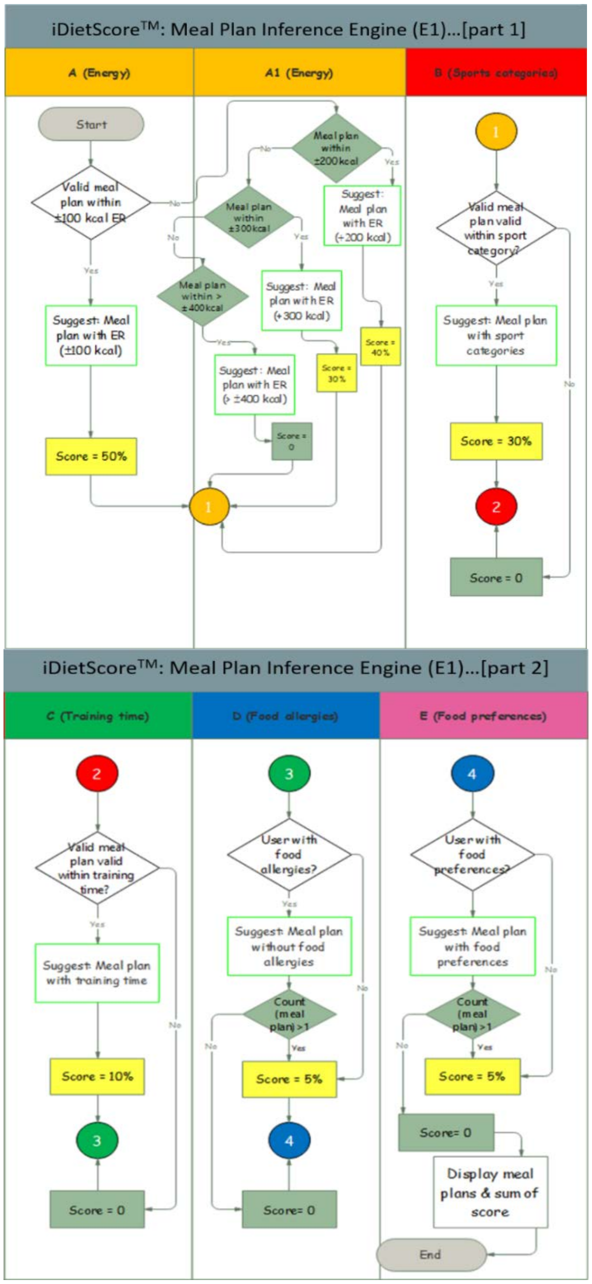

iDietScore™: Meal Plan Inference Engine (E1)...[part 1]
A (Energy)
A1 (Energy)
B (Sports categories)
Start
Valid meal plan within ±100 kcal ER
No
Yes
Suggest: Meal plan with ER (±100 kcal)
Score = 50%
Meal plan within ±200kcal
No
Yes
Suggest: Meal plan with ER (+200 kcal)
Meal plan within ±300kcal
No
Yes
Suggest: Meal plan with ER (+300 kcal)
Meal plan within > ±400kcal
Yes
Suggest: Meal plan with ER (> ±400 kcal)
Score = 0
Score = 30%
Score = 40%
1
Valid meal plan valid within sport category?
Yes
No
Suggest: Meal plan with sport categories
Score = 30%
2
Score = 0
iDietScore™: Meal Plan Inference Engine (E1)...[part 2]
C (Training time)
D (Food allergies)
E (Food preferences)
2
Valid meal plan valid within training time?
Yes
No
Suggest: Meal plan with training time
Score = 10%
3
Score = 0
3
User with food allergies?
Yes
No
Suggest: Meal plan without food allergies
Count (meal plan) >1
Yes
No
Score = 5%
4
Score= 0
4
User with food preferences?
No
Suggest: Meal plan with food preferences
Count (meal plan) >1
Yes
No
Score = 5%
Score= 0
Display meal plans & sum of score
End

Fig. 4. A the Flow of Meal Plan Inference Engine (E1).

Based on the system architecture, as the user received the meal plan list, they must choose one that they preferred. Next, as they agreed with all the food that is suggested in the meal plan, they must confirm it, and that particular meal plan is called an individualized meal plan. However, they have an option to change any specific food that they disagreed within the meal plan. Thus, inference engine 2 (E2) will infer a suitable food for changes based on meal reconstruction rules. Three rules must be met for meal reconstruction that included in the knowledge base. The rules are labeled with alphabetic and the details are described in Table II. Rule A related to the food group meanwhile rule B is about the calories of current food. In rule B, the calorie allowance is depending on the food group. A new food group (Food group A – Food group D) was created to classify them based on calorie allowance. The aim is to make sure the food list will not be exceeding or lacking any macronutrients and at the same time able to provide varieties of food choices to the user. The concept is to include a user's preference in planning on their meal plan within the control environment. The process flow (Fig. 5) shows how the rules are being applied in the inference engine 2 (E2) for food changes in meal reconstruction. The meal reconstruction aims to include user participation in developing their meal plan by allowing them to choose food that they preferred while being controlled by specific rules. The applied rules assure the calories and macronutrient is within the recommendation.

E. Database

The availability of the local food and beverages database is a critical component that can contribute to mobile app user satisfaction, especially in the food record feature. This study modified the food database established by [39] to develop a sports nutrition mobile app. The current study added another 2000 common Malaysian multi-ingredient menus in the database selected from the National Sports Council of Malaysia cafeteria and sports school dining hall. The food items were label into 32 food groups, six food types,13 food exchange groups and seven mealtimes (breakfast, lunch, dinner, snacks, pre-exercise meal, during and post-exercise meal). The food database contains the information on energy (kcal), house-hold measurement unit, eight nutrient contents (in gram) of over 3000 food items that are commonly eaten and prepared in Malaysian.

Meal plan databases were compiled using iDietScore^{TM} web app. SD/SN from ISN was voluntary use the web and develop a meal plan. In order to make sure the meal plan database can cover most of the athlete's profile, SD/SN create a meal planning based on calories (of 1500, 2500, 3500 and 4500 kcal), sports categories (of endurance, intermittent/power strength, skill and active individuals), season (of "on season" and "off-season") and training time (morning training, evening training and morning and evening training). There was 120 meal plan database were collected and used in the expert system. The meal plan number is expected to be increased as the web is released and used by more SD/SN.

TABLE II. THE THREE RULES OF MEAL RECONSTRUCTION

	Rules description	If ..then..	Example
A	Food group	IF current food group is A THEN list out food that within food group A	If user want to change apple (from food group named fruits), the food list must come from fruits.
B	Calorie: ±100 (total calorie)*	IF current calorie is B THEN list out food that within calories ±10 of B not exceed total 100 kcal/day	If apple calories is 75 kcal, the list of fruits must within ±10 of 75kcal
C	Mealtime (Breakfast, Snack, Lunch, Dinner, pre-exercise, during and post-exercise)	IF current mealtime is C THEN list out food that tag with mealtime C	If user want to change food for breakfast, the list of new food must also suitable for breakfast.

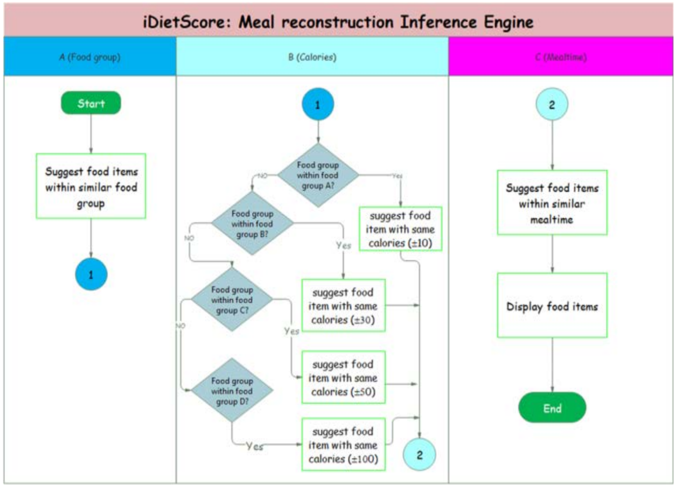
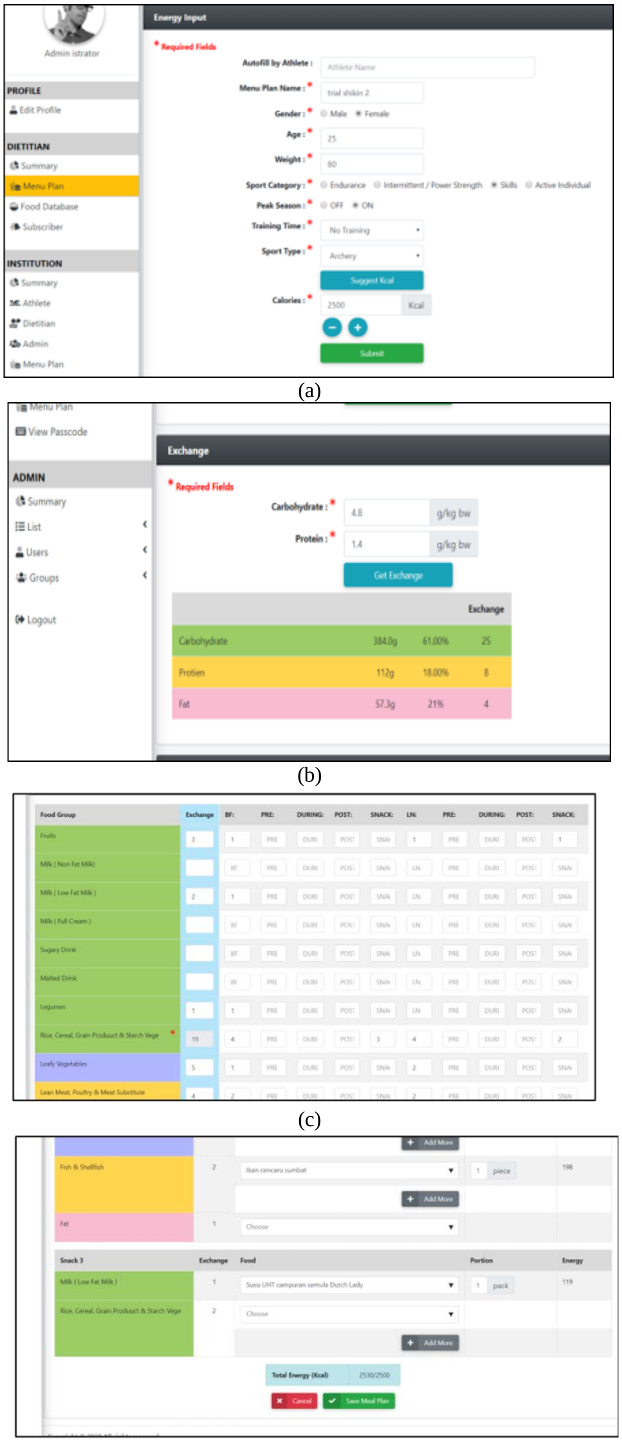

iDietScore: Meal reconstruction Inference Engine
A (Food group)
B (Calories)
C (Mealtime)
Start
Suggest food items within similar food group
1
1
Food group within food group A?
NO
Yes
Food group within food group B?
NO
Yes
suggest food item with same calories (±10)
Food group within food group C?
NO
Yes
suggest food item with same calories (±30)
Food group within food group D?
Yes
suggest food item with same calories (±50)
suggest food item with same calories (±100)
2
2
Suggest food items within similar mealtime
Display food items
End

Fig. 5. The Flow of Meal Reconstruction Inference Engine (IE 2).

F. Development and Expert Evaluation of a Prototype

Agile methodologies, an iterative process of development, were applied in this study. Agile development promotes adaptive planning, evolutionary development, early delivery, continuous improvement, and encourages rapid and flexible response to change [40]. This approach makes the developers more aware of the changing nature of the project and at the same time, also lets them give feedback to the design process [41]. The agile process will force all the members to collaborate and coordinate among team members during the development process [41]. Thus, the integration of agile methodologies is believed to improve the communication between two different fields of knowledge, such as health and technology. Fig. 6 shows the design of important features in iDietScore^{TM} web for SD/SN. Fig. 7 shows the design of essential features in iDietScore^{TM} mobile app for athletes and active individualized.

The expert evaluation was conducted upon the completion of both prototypes. Eight expert panels (in sports nutrition area) were involved. They need to evaluate the relevancy of features content, appropriateness, easiness, and accuracy of the language used and the attractiveness and appropriateness of the graphic user interface (GUI) based on the four-point scale. Scale one indicates not relevant/appropriate/easy/accurate/attractive. Scale two means the item needs revision, scale three is relevant but needs minor revision and scale four is very relevant/appropriate/ easy/accurate/attractive. All experts (8/8) agreed that all the features content was relevant to the target population (55% of experts rated the features as very relevant and the remaining rated the features as relevant but need minor revision). Moreover, 75% (6/8) of experts rated the language used in web and mobile app were appropriate, easy to understand and accurate. Lastly, 62.5% (5/8) viewed that the GUI was appropriate and attractive.



* Required Fields

Autofill by Athlete : Athlete Name

Menu Plan Name : * trial shikin 2

Gender : * ○ Male ● Female

Age : * 25

Weight : * 80

Sport Category : * ○ Endurance ○ Intermittent / Power Strength ● Skills ○ Active Individual

Peak Season : * ○ OFF ● ON

Training Time : * No Training

Sport Type : * Archery

Calories : * 2500 Kcal

(a)

* Required Fields

Carbohydrate : * 4.8 g/kg bw

Protein : * 1.4 g/kg bw

			Exchange
Carbohydrate	384.0g	61.00%	25
Protien	112g	18.00%	8
Fat	57.3g	21%	4

(b)

Food Group	Exchange	BF:	PRE:	DURING:	POST:	SNACK:	LN:	PRE:	DURING:	POST:	SNACK:
Fruits	3	1	PRE	DURI	POS	SNA	1	PRE	DURI	POS	1
Milk (Non Fat Milk)		BF	PRE	DURI	POS	SNA	LN	PRE	DURI	POS	SNA
Milk (Low Fat Milk)	2	1	PRE	DURI	POS	SNA	LN	PRE	DURI	POS	SNA
Milk (Full Cream)		BF	PRE	DURI	POS	SNA	LN	PRE	DURI	POS	SNA
Sugary Drink		BF	PRE	DURI	POS	SNA	LN	PRE	DURI	POS	SNA
Malted Drink		BF	PRE	DURI	POS	SNA	LN	PRE	DURI	POS	SNA
Legumes	1	1	PRE	DURI	POS	SNA	LN	PRE	DURI	POS	SNA
Rice, Cereal, Grain Produuct & Starch Vege *	19	4	PRE	DURI	POS	3	4	PRE	DURI	POS	2
Leafy Vegetables	5	1	PRE	DURI	POS	SNA	2	PRE	DURI	POS	SNA
Lean Meat, Poultry & Meat Substitute	4	2	PRE	DURI	POS	SNA	2	PRE	DURI	POS	SNA

(c)

		+ Add More		
Fish & Shellfish	2	Ikan cencaru sumbat ▼	1 piece	198
		+ Add More		
Fat	1	Choose ▼		
Snack 3	Exchange	Food	Portion	Energy
Milk (Low Fat Milk)	1	Susu UHT campuran semula Dutch Lady ▼	1 pack	119
Rice, Cereal, Grain Produuct & Starch Vege	2	Choose ▼		
		+ Add More		

Total Energy (Kcal) 2530/2500

(d)

Fig. 6. Snapshot of Important Features in iDietScore^{TM} Web for Sports Dietitians or Nutritionist: (a) Enter user's Profile; (b)Determine Energy and Macronutrient (c) Generate Exchange Table by Food Groups and Mealtime; (d) Create Meal Plan Meal by suggesting Suitable Food Items based on Exchange Distribution.

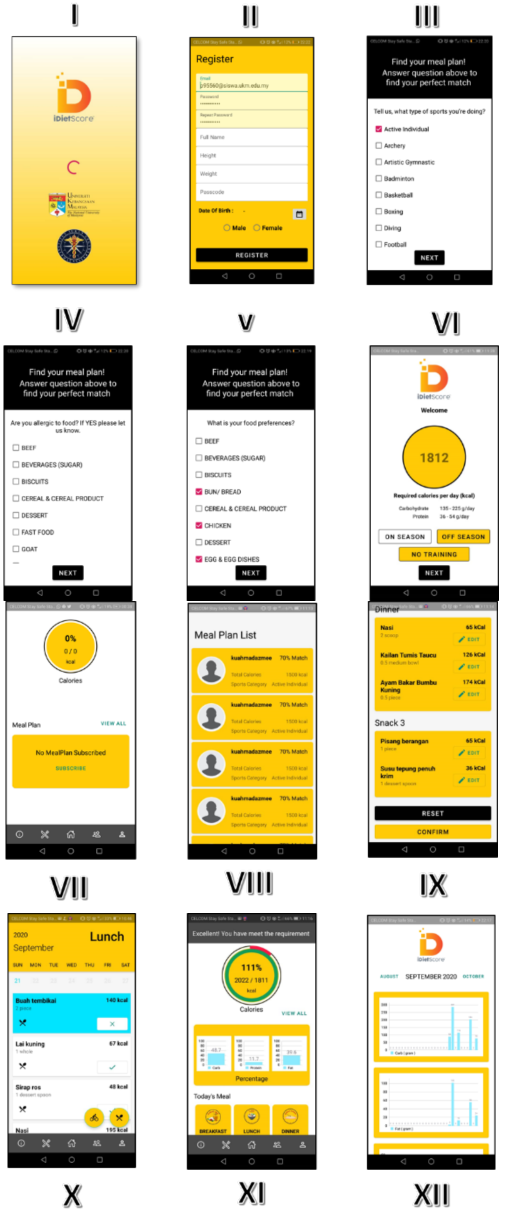


Fig. 7. Snapshot of Main Features in iDietScore^{TM} Mobile App for Athletes and Active Individuals: (I-VI) Athletes Profile that Include Height, Weight, Sports Type, Food Allergic, Food Preferences, Seasons and Training Time; (VII) Subscribe Meal Plan; (VIII) List of Meal Plan that Suits Athletes Profile; (IX) Meal Reconstruction; (X) Food Log; (XI) Home Page with Current Energy and Macronutrient Intake; (XII) Analysis for 1 month.

IV. Conclusion and Future Work

In conclusion, the design of iDietScore^{TM} was developed to provide a virtual meal plan based on nutrition need, training plan, training cycle (the season) and food preferences for athletes and active individuals. The rule-base expert system consisting of an inference engine matches a user profile with a meal plan database through meal plan rules integrated into the knowledge base. Moreover, the meal reconstruction rule set in the iDietScoreTM expert system's knowledge base allowed the user to change any suggested food items in the original meal plan into a food item that they preferred without altering their macronutrient needs. Users' participation in creating meal plans is another key element of the individualized meal plan, ultimately creating a sense of control and motivation to help individuals maintain short- and long-term adherence.

The future study would be the usability test (formative and summative testing) among target users. The formative testing aims to reveal several usability related problem that faces by the target user. These problems can be fixed or solve before final development is deployed. The think-aloud method was commonly used captured user experience of using the prototype. Meanwhile, the summative focus on the efficacy of the final system development and to obtain definitive evidence of usability on the final product. In addition, two experts in sports nutrition will independently evaluate the suitability of meal planning recommended by the inference engine based on 30 case studies that mimic the user's profile. The interrater agreement can be measured to make sure the rating obtained are not distinctive results of one rater's subjective judgment.

Acknowledgment

The authors want to thank the National Sports Institute Malaysia and Universiti Kebangsaan Malaysia for funding and supporting this research under grant code: NN-2018-093.

References

[1] D. T. Thomas, K. A. Erdman, and L. M. Burke, "Position of the Academy of Nutrition and Dietetics, Dietitians of Canada, and the American College of Sports Medicine: Nutrition and Athletic Performance," J. Acad. Nutr. Diet., vol. 116, no. 3, pp. 501–528, 2016.

[2] T. Stellingwerff, J. P. Morton, and L. M. Burke, "A framework for periodized nutrition for athletics," Int. J. Sport Nutr. Exerc. Metab., vol. 29, no. 2, pp. 141–151, 2019.

[3] P. Tumnark, J. P. Vilas-boas, P. Cardoso, and J. Cabral, "Ontology-Based Personalized Dietary Recommendation for Weightlifting Ontology-Based Personalized Dietary Recommendation for Weightlifting," Int. Work. Comput. Sci. Sport., no. December, 2014.

[4] J. Dwyer, A. Eisenberg, K. Prelack, W. O. Song, K. Sonneville, and P. Ziegler, "Eating attitudes and food intakes of elite adolescent female figure skaters : a cross sectional study," pp. 1–7, 2012.

[5] A. R. Jagim, H. Zabriskie, B. Currier, P. S. Harty, R. Stecker, and C. M. Kerksick, "Nutrient Status and perceptions of energy and macronutrient intake in a Group of Collegiate Female Lacrosse Athletes," vol. 7, pp. 1–7, 2019.

[6] S. L. Jenner, G. L. Buckley, R. Belski, B. L. Devlin, and A. K. Forsyth, "Team Sport Athletes Do Not Meet Sport Nutrition Recommendations — A Systematic Literature Review," no. 3, pp. 1–16, 2019.

[7] S. L. Jenner et al., "Dietary intake of professional Australian football athletes surrounding body composition assessment," J. Int. Soc. Sports Nutr., vol. 15, no. 1, pp. 1–8, 2018.

[8] M. Noll, C. R. De Mendonça, L. P. De Souza Rosa, and E. A. Silveira, "Determinants of eating patterns and nutrient intake among adolescent athletes: A systematic review," Nutr. J., vol. 16, no. 1, pp. 1–11, 2017.

[9] J. M. Spahn et al., "State of the Evidence Regarding Behavior Change Theories and Strategies in Nutrition Counseling to Facilitate Health and Food Behavior Change," J. Am. Diet. Assoc., vol. 110, no. 6, pp. 879–891, 2010.

[10] C. J. Boushey, A. M. Coulston, C. L. Rock, and E. Monsen, Nutrition in the prevention and treatment of disease. Academic Press, 2001.

[11] J. G. Mirtschin et al., "Organization of Dietary Control for Nutrition-Training Intervention Involving Periodized Carbohydrate Availability and Ketogenic Low-Carbohydrate High-Fat Diet," Int. J. Sport Nutr. Exerc. Metab., vol. 28, no. 5, pp. 480–489, Sep. 2018.

[12] W. K. Chen, The electrical engineering handbook. United States of America: Elsevier Academic Press, 2004.

[13] S. Alian, J. Li, and V. Pandey, "A Personalized Recommendation System to Support Diabetes Self-Management for American Indians," Inst. Electr. Electron. Eng., vol. 6, pp. 73041–73051, 2018.

[14] M. Burgermaster et al., "A new approach to integrating patient-generated data with expert knowledge for personalized goal setting: A pilot study," Int. J. Med. Inform., vol. 139, no. February, p. 104158, 2020.

[15] C. H. Chen, M. Karvela, M. Sohbati, T. Shinawatra, and C. Toumazou, "PERSON - Personalized Expert Recommendation System for Optimized Nutrition," IEEE Trans. Biomed. Circuits Syst., vol. 12, no. 1, pp. 151–160, 2018.

[16] I. Marinchev and G. Agre, "An expert system for healthful and dietary nutrition," ACM Int. Conf. Proceeding Ser., vol. 1164, no. June, pp. 229–236, 2016.

[17] S. A. Noah et al., "DietPal: A Web-based dietary menu-generating and management system," J. Med. Internet Res., vol. 6, no. 1, pp. 32–48, 2004.

[18] G. Jaswal, A. Kaul, and R. Nath, "Knuckle print biometrics and fusion schemes - Overview, challenges, and solutions," ACM Comput. Surv., vol. 49, no. 2, 2016.

[19] H. Jung and K. Chung, "Knowledge-based dietary nutrition recommendation for obese management," Inf. Technol. Manag., vol. 17, no. 1, pp. 29–42, 2016.

[20] C. L. Bush et al., "Toward the Definition of Personalized Nutrition: A Proposal by The American Nutrition Association," J. Am. Coll. Nutr., vol. 39, no. 1, pp. 5–15, 2020.

[21] A. T. Limon-miro, V. Lopez-teros, and H. Astiazaran-garcia, "Dynamic Macronutrient Meal-Equivalent Menu Method : Towards Individual Nutrition Intervention Programs," methods Protoc., vol. 2, no. 3, p. 78, 2019.

[22] A. Kale and N. Auti, "Automated menu planning algorithm for children: food recommendation by dietary management system using ID3 for Indian food database," in Procedia Computer Science, 2015, vol. 50, pp. 197–202.

[23] L. Yang et al., "Yum-Me: A personalized nutrient-based meal recommender system," ACM Trans. Inf. Syst., vol. 36, no. 1, 2017.

[24] R. Yera Toledo, A. A. Alzahrani, and L. Martinez, "A food recommender system considering nutritional information and user preferences," IEEE Access, vol. 7, pp. 96695–96711, 2019.

[25] I. M. Ahmed, M. Alfonse, and M. A. A. M. Salem, "Daily Meal Planner Expert System for Diabetics Type-2," E-Leader Prague, no. 2007, 2015.

[26] M. V Gupta, P. Bhattacharjee, and N. Kotian, "DANES: Diet and Nutrition Expert System for Meal Management and Nutrition Counseling," Int. J. Recent Innov. Trends Comput. Commun. , vol. 5, no. 12, pp. 204–208, 2017.

[27] B. A. Al-dhuhli, "Developing a Nutrition and Diet Expert System Prototype," no. JUNE 2013, pp. 1368–1375, 2013.

[28] A. H. Mohammad and N. A. M. Al Saiyd, "A Framework For Knowledge Acquisition," Int. J. Comput. Sci. Netw. Secur., vol. 10, no. May, pp. 42–50, 2010.

[29] V. Espín, M. V. Hurtado, and M. Noguera, “Nutrition for Elder Care: A nutritional semantic recommender system for the elderly,” Expert Syst., vol. 33, no. 2, pp. 201–210, 2016.

[30] D. Ribeiro, J. Machado, J. Ribeiro, M. J. M. Vasconcelos, E. F. Vieira, and A. Correia de Barros, “SousChef: Mobile Meal Recommender System for Older Adults,” no. Ict4awe, pp. 36–45, 2017.

[31] F. Mata, M. Torres-Ruiz, R. Zagal, G. Guzman, M. Moreno-Ibarra, and R. Quintero, "A cross-domain framework for designing healthcare mobile applications mining social networks to generate recommendations of training and nutrition planning," Telemat. Informatics, vol. 35, no. 4, pp. 837–853, 2018.

[32] A. Pilnick and J. A. Swift, "Qualitative research in nutrition and dietetics: Assessing quality," J. Hum. Nutr. Diet., vol. 24, no. 3, pp. 209–214, 2011.

[33] C. Anderson, "Presenting and Evaluating Qualitative Research: Strengths and Limitations of Qualitative Research," Am. J. Pharm. Educ., vol. 74, no. 8, pp. 1–7, 2010.

[34] A. Draper and J. A. Swift, "Qualitative research in nutrition and dietetics: Data collection issues," J. Hum. Nutr. Diet., vol. 24, no. 1, pp. 3–12, 2011.

[35] V. Braun and V. Clarke, "Using thematic analysis in psychology," Qual. Res. Psychol., vol. 3, no. 2, pp. 77–101, Jan. 2006.

[36] C. M. Kerksick et al., "ISSN exercise & sports nutrition review update: Research & recommendations," J. Int. Soc. Sports Nutr., vol. 15, no. 1, pp. 1–57, 2018.

[37] S. Shahar, N. S. Safii, Z. Manaf, and H. Haron, Atlas Makanan: Saiz Pertukaran & Porsi, 3rd editio. Kuala lumpur: MDC Publishers Sdn Bhd, 2015.

[38] N. N. R. Rodriguez, N. Di Marco, S. Langley, and N. M. DiMarco, "American College of Sports Medicine, American Dietetic Association, and Dietitians of Canada joint position statement: Nutrition and athletic performance.," Med. Sci. Sports Exerc., vol. 41, no. 3, pp. 709–731, 2009.

[39] S. Nik Shanita et al., "DietScoreTM: Sports Nutrition-based Mobile Application for Athletes and Active Individuals," in IFMBE Proceedings, vol. 58, 2017, pp. 1–5.

[40] C. Quesada-lópez et al., "Design , Development and Validation of a Mobile Application for Goal Setting and Self-Monitoring of Dietary Behaviors," no. 2004, 2009.

[41] E. A. Altameem, "Impact of Agile Methodology on Software Development," vol. 8, no. 2, pp. 9–14, 2015.

Learner Behavior in e-Learning as a Multicriteria Attribute based on Perspective of Flow Experience

Dadang Syarif Sihabudin Sahid
Department of Information Technology
Politeknik Caltex Riau, Pekanbaru, Indonesia

***Abstract*—Flow experience describe psychological condition in the form of optimal experience of an activity. Flow shows an interconnection, interest, and pleasure toward an activity, thus enable user to fully participate in the activity. In e-learning activity, flow provides positive experience of a learning process. This condition is essential in user's ability to achieve high performance. Therefore, it is important to identify user's flow experience during his interaction with e-learning. This information can be used as reference of how e-learning model provide response that in accordance with user's psychological condition. Assessment of psychological experience based on flow theory have been conducted in many studies, particulary based on experience sampling method. However, these survey methods require high effort thus they are inefficient. The previous studies in this topic only covers conventional learning, with face-to-face interaction. In e-learning, particulary those that use adaptive context aware e-learning approach, flow experience can be assessed by conducting inference based on learning behavior parameters of learners during interaction with e-learning. However, there is no study that provide relation among learner's learning behavior in e-learning with parameters of flow experience. Therefore, this study tested hypotheses aimed to obtain relation between learning behavior and flow experience. Hypotheses model constructed by involving technology acceptance model (TAM), expectation confirmation model, and flow experience as learning psychological condition. Learning behavior as a multicriteria attribute was represented by actual usage in form of intensity of using e-learning. Meanwhile, perceived balance of skill and challenge as representation of flow experience was selected as main variable in the proposed hypotheses. The result showed that these variables had positive relation with each other.**

***Keywords*—Flow experience; learning behavior; multicriteria attribute; TAM; e-learning**

I. Introduction

Along with the development of ubiquitous and pervasive computing, e-learning technology adapts with users' requirement and behavior [1]. E-learning, which initially follows teacher centered learning design, has developed into student centered learning at present. This means that e-learning has been developed with emphasize to adaptive aspects and user personalization. In the past, e-learning was a rigid system with one procedural line in providing knowledge to learners. Meanwhile, e-learning nowadays provides adaptive learning to satisfy learners' requirement, situation, and behavior. One way to achieve adaptive e-learning is by developing context aware e-learning.

The implementation of context-aware in e-learning system motivates learners to be actively engaged in learning process. Context-aware e-learning provides adaptive learning that satisfies learners' requirement, situation, and behavior. This system is able to overcome important issues in e-learning implementation, which one of them is how e-learning system considers learners' psychological condition during interaction with the system.

As presented by Zhang and Cheng [2], the development of context aware e-learning personalization model with consideration of learners' psychological condition as context and reference to specific learning pattern is limited. Yet, Whitson and Consoli [3] stated that the learners' psychological condition plays important role in motivating learners in learning process.

The most common psychological conditions experienced by learners during learning is anxiety, boredom, and optimal. These three conditions follow the flow theory [4], that analyzes motivation, emotional, and cognitive condition of learners [5]. Flow state is proposed by Csikszentmihalyi [6], describing condition when learners' skills (experience, knowledge, cognitive skill) are proportional to the given challenges/problems. Meanwhile, anxiety occurs when learner's skills are lower than the level of given challenges/problems. In contrast, boredom occurs when learners skills are higher than the level of given challenges/problems.

Studies of psychological conditions identification based on flow theory generally conducted in conventional learning process as e-learning is not supported by context aware personalization technology. Studies of psychological experience assessment also often conducted via surveys and questionnaires [7] that spent long time and require high effort. However, as in conventional learning process, learning behavior in e-learning can assess learners' psychological condition during interaction by their behaviors and characteristic [8]. Learning behavior can be used to assess one person acceptance toward a technology, especially e-learning technology. Positive learning behavior will induce positive psychological experience shown by learner's active engagement toward e-learning. Hence, this study observed relation among learning behavior as multicriteria attribute, technology acceptance based on TAM variables, and flow experience components as a representation of learning psychological condition. This paper is organized as follows: Introduction (Section I), Literature Review (Section II), Hypotheses Development (Section III), Method (Section IV),

This research is sponsored by Ministry of Research, Technology, and Higher Education, Republic of Indonesia, 2020.

Results and Discussion (Section V), and Conclusion (Section VI).

II. Literature Review

The literature review is focused on components related to technology acceptance model (TAM), flow experience, and learning intensity as representation of learning behavior. These components are perceived ease of use, perceived usefulness, attitude, intention, enjoyment, perceived control, skill-challenge balance, engagement and learning intensity.

A. Perceived Ease of Use

Perceived ease of use is a level of how much one person can trust an information technology to be able to easily used [9]. Intensity of use and interaction between user and system also can be considered as indicators of ease of use [10]. When a system is frequently used, it means that the system is easy to understand, operate, and implement. Perceived ease of use [11] is an individual trust level toward whether the technology can aid him in completing his task.

Based on the aforementioned explanation, ease of use of an information technology depended on individual trust level toward whether the system is easy to understand, operate, implement. Therefore, common indicators of perceived ease of use are: information technology is easy to understand, information technology is easy to implement, and information technology is easy to operate.

B. Perceived Usefulness

Perceived usefulness is defined as an individual trust level toward how the use of a subject benefits the user [9][10]. These benefits can be considered from various point of views [12], such as:

- Makes job easier
- Useful
- Increase productivity
- Enhance effectiveness
- Improve job performance

C. Attitude

Attitude toward e-learning is user evaluation, either positive or negative one, of e-learning use. This evaluation is related to favorable or unfavorable behavior toward e-learning use [13]. This attitude is an internal state that influenced individual thought of specific object, person, or activity. Attitude is a cognitive, affective, and behavior tendency that is studied to give positive or negative response toward specific object, situation, institution, concept, or person [14]. Attitude is a personal factor that contains positive evaluation or inner behavior of avoiding, resisting, or blocking an object [15].

D. Intention

Intention in TAM is a behavior tendency toward use of technology [9]. Intention is user psychic aspect that tends to pay attention of prefer the use of technology to meet his objective [16]. Intention to use can be defined as user intention to use or reuse a specific technology.

Intention to use can be assessed by at least two factors: Compatibility, which is a level where user perceives an innovation based on his value, past experience, and potential requirement; Ease of use, which is a level where user perceives a specific subject is easy to the implement or operate.

E. Perceived Control

Perceived control is a person ability to control an event, subject, technology or other conditions [17]. This condition is related with internal control of situation or belief to influence other person or the environment. A person with high perceived control tends to be highly motivated in meeting his objective, have potential and necessary skill to overcome a situation. In contrast, a person with low perceived control tends to have a passive attitude [18][19].

F. Enjoyment

In several literatures, enjoyment is often associated as pleasure, interest, comfort, and other positive feeling. In learning, enjoyment is something frequently conducted that relates to mastering a specific skill [6]. Enjoyment is a level where activity of using a technology brings pleasure to the user, apart from the value of the technology itself. Enjoyment can also be explained as a high comfort of a person during interaction with information technology. If perceived usefulness is regarded as external motivation, then perceived enjoyment is an internal motivation of information technology use [11].

G. Learning Intensity

Learning intensity in e-learning represents one person's behavior during interaction with e-learning (learning behavior). As with conventional learning, intensity of attending classes, intensity of doing assignments, activity of collaboration, and activity of exploration are indicators of whether one person engaged in learning. In e-learning situation, this learning behavior can be observed, noted, and explored from log files in server [20]. Several observed behaviors are: learning participation, learning duration, duration to complete assignments, and number of completed assignments. These are examples of exploration behavior of learning material.

H. Balance of Skill and Challenge

Balance of skill and challenge is often referred to assess flow psychology experience occurrence. Flow theory is firstly explained by Mihaly Csikszentmihalyi who use the word 'flow' to represent optimal experience of someone to be focused to his involvement in an activity [6]. An individual in flow condition is in state of high concentration so that there are no space for other thoughts or disturbance. Even though flow is constructed of various complex variables, skill and challenge are two of the most important variables [21][22].

I. Student Engagement

Student engagement is essential in learning process, particularly in e-learning. One of the challenges is building students' interest on learning. Engagement in learning will bring positive effect on student learning performance. Something noticeable such as student's behavior when accessing e-learning, but also something unnoticeable such as psychological condition. This condition related to persistence and consistency level of a student and his positive emotion during interaction with e-learning.

Fredricks et al. [23] states that engagement is a combination of three dimensions: behavioral, cognitive, and emotional. Behavioral engagement relates with students' behaviors observed during interaction in learning process, such as attendance, activeness in completing assignments, and participation in learning process [24]. Cognitive engagement generally includes mastering of skill, knowledge, idea [25], internal motivation to learn something, and ability to plan, supervise, and evaluate a specific learning. Emotional engagement relates with students' feeling and emotion during learning, such as excitement, boredom, anxious, sadness, and other emotions [26].

III. Hypotheses Development

Hypotheses in this study was constructed by involving components of flow theory, extension of expectation confirmation model, and technology acceptance model (TAM) in learning with application of e-learning. Flow theory and TAM has interconnected components, for example component of perceived ease of use in TAM is interconnected with component of skill and challenge in flow theory. Other interconnected components are perceived usefulness, attitude, intention, and actual usage.

As stated in previous literature review, skill and challenge have positive effect on perceived ease of use and perceived usefulness. In TAM extension, skill-challenge is shows external variable of relevant skill and challenge [11]. Someone with adequate skill of prior knowledge and web experience, in addition to being able to overcome given challenge will bring positive influence on perceived ease of use and perceived of usefulness. Skill-challenge balance illustrates the occurrence of flow experience. Hence, the first, second, third and fourth hypotheses of this study were:

- H1: Skill and challenge balance brings positive influence on perceived ease of use.
- H2: Skill and challenge balance brings positive influence on perceived usefulness.
- H3: Skill and challenge balance brings positive influence on enjoyment.
- H4: Skill and challenge balance brings positive influence on perceived control.

Student with perceived ease of use can navigate e-learning use easily. Perception also affect perceived usefulness of e-learning. This is a valid relation as explained by previous studies involving TAM model [9]. This also valid for relation of perceived usefulness, attitude, intention and actual usage.

As explained before in student engagement, learning intensity is a representation of user behavior during interaction with e-learning. This component is something that can be used to illustrate user characteristic when using e-learning. In TAM model, learning intensity can be observed as part of user actual usage when using e-learning, which can be observed and measured quantitatively. A person with perceived usefulness toward e-learning will have attitude that displays e-learning has positive influence on learning. This attitude is followed with intention that yielded proper intensity when using e-learning. Based on this information, hypotheses fifth to ninth can be stated as:

- H5: Perceived ease of use bring positive influence on perceived usefulness.
- H6: Perceived usefulness bring positive influence on attitude.
- H7: Perceived ease of use bring positive influence on attitude.
- H8: Attitude bring positive influence on intention.
- H9: Attitude bring positive influence on actual usage.

Ming-Chi Lee study proves that there is a relation among flow parameters of enjoyment, concentration, and perceived control with attitude and intention [27]. From the previous relation, intention has an influence on actual usage that can be observed when user use e-learning persistently and consistently. This consistency is a mark of learning intensity. Therefore, the tenth, eleventh, and twelfth hypotheses can be stated as:

- H10: Enjoyment brings positive influence on attitude.
- H11: Perceived control brings positive influence on attitude.
- H12: Perceived control brings positive influence on intention.

These hypotheses create an outer model as displayed in Fig. 1.

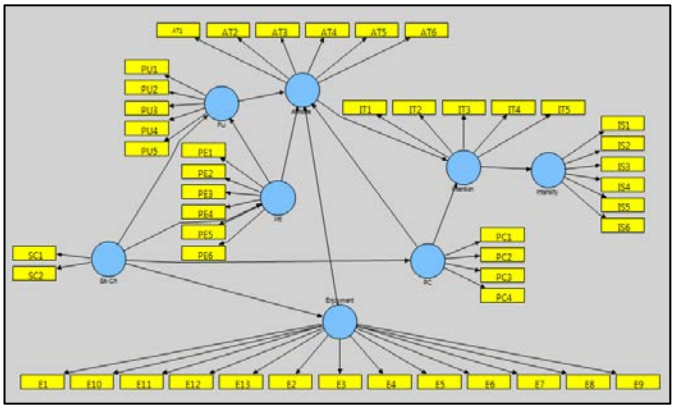


Fig. 1. Outer Model.

IV. Method

This was an explanatory study that explained causal relation among study variables by testing hypotheses with survey method. This method is suitable for testing study hypotheses, whether it is a descriptive, comparative, or associative hypotheses.

A. Constructs and Indicators

This study was approached with combination of flow theory and TAM components relating the use of e-learning. As illustrated in previous structure model, constructs and indicators of this study are shown at Table I.

TABLE I. CONSTRUCTS AND INDICATORS

No.	Construct	Indicator
1.	Perceived Balance Skill-Challenge	SC1. Mastering of skill SC2. Mastering of challenge
2.	Perceived Ease of Use (PE)	PE1. Ease of use PE2. Ease of study PE3. Ease of navigation PE4. Flexibility of use PE5. Ease of finding material PE6. Stability and continuity of use
3	Perceived usefulness (PU)	PU1. Ease of learning process PU2. Effectiveness of learning PU3. Skill improvement PU4. Ease of learning activity PU5. Benefit of learning
4	Learning Intensity	IS1. Frequency of use IS2. Duration of use IS3. Frequency of interest IS4. Duration of interest IS5. Frequency of study achievement IS6. Duration of study achievement
	Enjoyment	E1. Awakening (Drowsy – Awake) E2. Happiness (Sad – Happy) E3. Excited (Gloomy – Excited) E4. Strength (Weak – Strong) E5. Activeness (Passive – Active) E6. Individual relation (Lonely – Popular) E7. Self-confidence (Timid – Confident) E8. Involvement (Separated – Involved) E9. Interest (Bored – Interested) E10. Openness (Closed – Open) E11. Clarity (Confused – Clear) E12. Relaxation (Stressed – Relax) E13. Competitiveness (Competitive – Cooperative)
	Intention	IT1. Intention to use routinely IT2. Intention to use as main medium for learning IT3. Intention to use as communication event IT4. Intention to use as main medium of learning material IT5. Intention to use as main medium for practice
	Attitude	AT1. Using e-learning is good AT2. Using e-learning is important AT3. Using e-learning is interesting AT4. Using e-learning is positive AT5. Using e-learning is exciting AT6. Using e-learning is good for educational institution
	Perceived Control	PC1. Level of concentration PC2. Level of awareness when using e-learning PC3. Level of feeling when using e-learning PC4. Level of controlling situation

B. Data Collection

In this study, data was collected by conducting survey and direct observation in the field. Data was collected by interview, review of learning environment, review of learning facilities, review of distribution and geographic condition of participants, and providing questionnaire for learners. Interview was performed on teachers, headmaster and event organizer, and learners.

The questionnaire was based on constructs and indicators as illustrated in outer and inner models. Each questionnaire contained questions with alternative answers in attitude scale 1 to 5, which translated as ‘strongly disagree’ and ‘strongly agree’ attitudes, respectively. Several questionnaires also express level of negative and positive feeling toward use of e-learning.

The study objects were students of open high school in West Bandung Region, West Java. Open high school uses long distance learning model aided with information and communication technology. This model help disseminate and enlarge access for education for community limited with time, geographic environment, and socio-economic problems. The study population was 600 students, while the number of samples was 186 students of tenth, eleventh, and twelfth grade. The sample also selected by considering gender, employment, and distribution of geographic environment of study location.

C. Model Evaluation

The model was evaluated by observing variables of outer and inner models. Outer model was evaluated based on three criteria: convergent validity, discriminant validity, and reliability testing. Meanwhile, inner model was evaluated by using R-square analysis, path coefficient value, and observing the value of calculated t-value as reference of significance of construct variables relation.

Convergent validity with reflective indicator was assessed based on correlation between item score/component score and construct score, which was calculated with PLS. Reflective size is considered high if the correlation value is more than 0.70 to the calculated construct. Discriminant validity with reflective indicator was assessed based on cross loading of calculation with construct. If construct correlation with calculated item is larger than the size of other constructs, it means that latent construct predicts one block size better than the size of other blocks. Other methods used to find discriminant validity was comparison of square root of Average Variance Extracted (AVE) each construct with correlation value of one construct with the other (latent variable correlation). Reliability was evaluated by observing values of composite reliability and cronbach alpha. These value shows consistency of calculated indicators.

R-square value, path coefficient, and t-value of bootstrapping are commonly used to evaluate inner model. Interpretation of this value is similar with interpretation of conventional regression, which shows ability of independent variable to describe its dependent variable. Path coefficient is a standardized regression coefficient that shows direct effect of independent variable on a dependent variable inside the model. This coefficient shows direction (positive or negative) of an exogenous variable effect on endogenous variable. Meanwhile, t-value obtained from bootstrapping used to observe significance of resulting path coefficient. In this study, the t-value was compared with t-table with significance of 5%.

V. Results and Discussion

Based on data developed with SMART-PLS 2 application, the general results of this study are displayed in Fig. 2 and Fig. 3.

A. Outer Model Evaluation

Outer model or calculated model was evaluated by variables of convergent validity, discriminant validity, and composite reliability. Convergent validity was evaluated by outer loadings value of all indicators in each construct. As can be seen in Table II and Table III, outer loading of all indicators were > 0.7, which means that reflective indicator used in this study was valid.

Discriminant validity of reflective indicator can be observed in cross loading between indicator and its construct. Cross loading output of PLS algorithm is displayed in Table IV. Cross loadings output shows that correlation of each indicator with its construct was higher than other constructs, which means that latent constructs predicted indicator in its own group better than indicator in other groups.

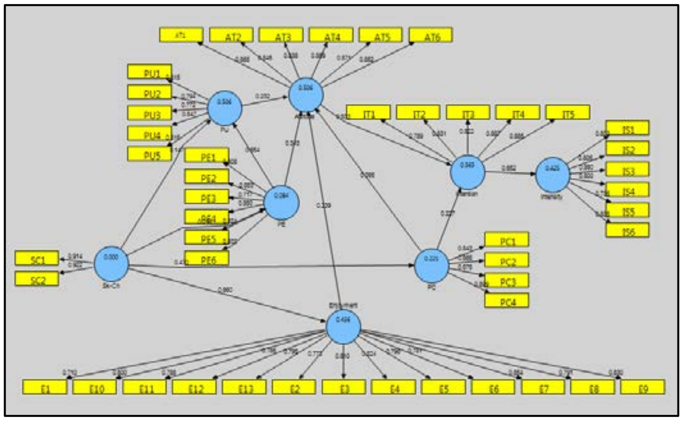


Fig. 2. Result of Developed Data using PLS.

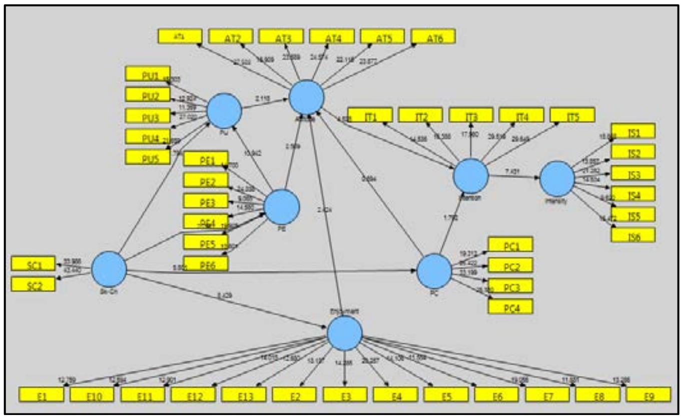


Fig. 3. Result of Developed Data using Bootstrapping.

TABLE II. Outer Loadings (1)

	Attitude	Enjoyment	Intensity	Intention
AT1	0.865060			
AT2	0.845336			
AT3	0.837621			
AT4	0.858831			
AT5	0.870568			
AT6	0.852354			
E1		0.709985		
E10		0.800327		
E11		0.788455		
E12		0.784546		
E13		0.795436		
E2		0.773233		
E3		0.809836		
E4		0.823621		
E5		0.798087		
E6		0.780855		
E7		0.853828		
E8		0.791450		
E9		0.829661		
IS1			0.803432	
IS2			0.805628	
IS3			0.860103	
IS4			0.800075	
IS5			0.735610	
IS6			0.834604	
IT1				0.789436
IT2				0.830746
IT3				0.821947
IT4				0.886985
IT5				0.884896

TABLE III. Outer Loadings (2)

	PC	PE	PU	Sk-C
PC1	0.842932			
PC2	0.866316			
PC3	0.877928			
PC4	0.898790			
PE1		0.808248		
PE2		0.853019		
PE3		0.717279		
PE4		0.850064		
PE5		0.823625		
PE6		0.801512		
PU1			0.815296	
PU2			0.793783	
PU3			0.771812	
PU4			0.847220	
PU5			0.816217	
SC1				0.913872
SC2				0.922269

TABLE IV. CROSS LOADINGS

	Atti tude	Enjoy ment	Inten sity	Inten tion	PC	PE	PU	Sk-Ch
AT1	**0.9**	0.5	0.7	0.7	0.3	0.6	0.5	0.4
AT2	**0.8**	0.5	0.6	0.6	0.3	0.5	0.6	0.3
AT3	**0.8**	0.4	0.7	0.6	0.4	0.6	0.5	0.3
AT4	**0.9**	0.5	0.6	0.6	0.4	0.6	0.5	0.4
AT5	**0.9**	0.4	0.7	0.7	0.4	0.6	0.5	0.3
AT6	**0.9**	0.4	0.6	0.6	0.3	0.4	0.5	0.3
E1	0.4	**0.7**	0.4	0.4	0.5	0.4	0.2	0.5
E10	0.4	**0.8**	0.5	0.4	0.4	0.3	0.3	0.5
E11	0.4	**0.8**	0.4	0.3	0.4	0.4	0.3	0.5
E12	0.3	**0.8**	0.3	0.4	0.5	0.3	0.4	0.6
E13	0.4	**0.8**	0.4	0.4	0.4	0.4	0.4	0.5
E2	0.4	**0.8**	0.4	0.5	0.5	0.3	0.3	0.6
E3	0.4	**0.8**	0.3	0.3	0.4	0.3	0.2	0.5
E4	0.4	**0.8**	0.5	0.4	0.5	0.4	0.4	0.5
E5	0.4	**0.8**	0.4	0.4	0.5	0.4	0.4	0.6
E6	0.5	**0.8**	0.4	0.4	0.4	0.4	0.4	0.4
E7	0.5	**0.9**	0.4	0.4	0.5	0.4	0.4	0.5
E8	0.5	**0.8**	0.4	0.4	0.5	0.4	0.4	0.5
E9	0.4	**0.8**	0.4	0.4	0.4	0.3	0.4	0.5
IS1	0.6	0.4	**0.8**	0.5	0.3	0.5	0.4	0.3
IS2	0.5	0.3	**0.8**	0.5	0.2	0.5	0.4	0.2
IS3	0.7	0.4	**0.9**	0.6	0.4	0.6	0.5	0.3
IS4	0.6	0.3	**0.8**	0.5	0.2	0.5	0.4	0.2
IS5	0.5	0.5	**0.7**	0.5	0.3	0.5	0.5	0.2
IS6	0.7	0.5	**0.8**	0.6	0.3	0.6	0.6	0.3
IT1	0.5	0.4	0.6	**0.8**	0.4	0.5	0.4	0.4
IT2	0.6	0.4	0.5	**0.8**	0.4	0.5	0.3	0.4
IT3	0.6	0.4	0.5	**0.8**	0.4	0.5	0.3	0.3
IT4	0.7	0.4	0.6	**0.9**	0.4	0.5	0.4	0.4
IT5	0.7	0.5	0.6	**0.9**	0.4	0.5	0.4	0.4
PC1	0.2	0.5	0.2	0.4	**0.8**	0.2	0.2	0.4
PC2	0.3	0.5	0.3	0.4	**0.9**	0.2	0.2	0.4
PC3	0.4	0.5	0.4	0.5	**0.9**	0.3	0.3	0.4
PC4	0.4	0.5	0.4	0.4	**0.9**	0.4	0.3	0.4
PE1	0.6	0.3	0.5	0.4	0.3	**0.8**	0.6	0.2
PE2	0.5	0.3	0.5	0.5	0.2	**0.9**	0.6	0.2
PE3	0.5	0.3	0.5	0.5	0.3	**0.7**	0.5	0.2
PE4	0.5	0.5	0.5	0.4	0.3	**0.9**	0.6	0.3
PE5	0.5	0.4	0.5	0.5	0.3	**0.8**	0.6	0.3
PE6	0.5	0.4	0.5	0.4	0.2	**0.8**	0.5	0.2
PU1	0.5	0.4	0.5	0.4	0.2	0.5	**0.8**	0.3
PU2	0.5	0.3	0.4	0.3	0.2	0.5	**0.8**	0.3
PU3	0.5	0.3	0.5	0.4	0.3	0.5	**0.8**	0.3
PU4	0.5	0.3	0.5	0.4	0.2	0.7	**0.8**	0.2
PU5	0.5	0.4	0.5	0.4	0.3	0.6	**0.8**	0.3
SC1	0.4	0.6	0.3	0.4	0.4	0.3	0.3	**0.9**
SC2	0.3	0.6	0.3	0.4	0.5	0.3	0.3	**0.9**

Other variables that can be used to evaluate discriminant validity is by comparison square root of AVE for each construct with correlation of one construct with others (Latent Variable Correlation).

Discriminant validity value is considered adequate if the square root of AVE for each construct is higher than latent variable correlation value. Output of AVE and latent variable correlation from PLS Algorithm are shown in Table V and Table VI. Based on comparison of both tables, the square root of AVE for each construct was higher than correlation of one construct with others.

Evaluation of construct reliability can be conducted based on two variables, composite reliability and croncbach alpha of indicator group that assess the construct. Construct is considered reliable if the composite reliability value and cronbach alpha are larger than 0.7. Output of composite reliability and cronbach alpha are displayed in Table VII.

TABLE V. AVE VALUE

	AVE	AVE Square root
Attitude	0.731086	0.855036
Enjoyment	0.633615	0.795999
Intensity	0.652021	0.807478
Intention	0.711746	0.843650
PC	0.759905	0.871725
PE	0.656464	0.810225
PU	0.654897	0.809257
Sk-Ch	0.842871	0.918080

TABLE VI. LATENT VARIABLE CORRELATION

	Atti tude	Enjoy ment	Inten sity	Inten tion	PC	PE	PU	Sk-Ch
Attitude	1							
Enjoyment	0.5	1						
Intensity	0.7	0.5	1					
Intention	0.7	0.5	0.7	1				
PC	0.4	0.6	0.4	0.5	1			
PE	0.6	0.4	0.6	0.6	0.4	1		
PU	0.6	0.4	0.6	0.5	0.3	0.7	1	
Sk-Ch	0.4	0.7	0.3	0.4	0.5	0.3	0.3	1

TABLE VII. COMPOSITE RELIABILITY AND CRONBACHS ALPHA

	Composite Reliability	Cronbachs Alpha
Attitude	0.942227	0.926444
Enjoyment	0.957345	0.951605
Intensity	0.918149	0.892857
Intention	0.92493	0.898253
PC	0.926757	0.894898
PE	0.919547	0.89453
PU	0.904573	0.868169
Sk-Ch	0.914735	0.813687

As can be seen in Table VII, the composite reliability value and cronbach alpha for each construct was larger than 0.70, which means that each construct was considered reliable.

B. Inner Model Evaluation

Inner model or structure model can be evaluated based on R-square, path coefficient, and t-value. R-square and path coefficient were obtained from PLS algorithm, while t-value generated from bootstrapping process. As previously explained, interpretation of R-square value is similar with interpretation of conventional regression. This value shows goodness fit of independent variable in describing its dependent variable. Table VIII shows the value of R-square.

Path coefficient and t-value are commonly used to evaluate model structure, and particularly used to test the study hypotheses. Path coefficient has positive and negative values used to test whether two variables related as described in the hypotheses. Meanwhile, t-value was used to test significance of relation shown in path coefficient. The value of path coefficient, t-value, p-value, and t-table with 0.05 significance 0.05 are shown in Table IX. Based on these data, it can be determined that nearly all hypotheses were accepted, except for relation between perceived control and attitude, which was insignificant. Generally, according to the results, learning behavior variables have positive relation significantly. However, it must pay attention during categorizing these variables. According to [28], different categorizing gave different results.

In the proposed inner model, there were two path that involved two mediation variables: (i) Path from perceived ease of use to attitude via mediation variable of perceived usefulness (PE→PU→AT); and (ii) path from perceived control to intention via mediation variable of attitude (PC→AT→IT). Using Sobel equation, both mediation variables were checked to determine whether they were significant or not (indirect effect). The relation of indirect effect with Sobel equation using standard normal distribution is displayed in Fig. 4.

If a and s_a are path coefficient and standard error of relation between X and Y, while b and s_b are path coefficient and standard error of relation between Y and Z, then standard normal distribution (z) of both relations is:

$$z = \frac{a * b}{\sqrt{a^2 * {s_b}^2 + b^2 * {s_a}^2}} \tag{1}$$

TABLE VIII. R-SQUARE

	R-Square
Attitude	0.505674
Enjoyment	0.435945
Intensity	0.424761
Intention	0.565391
PC	0.220707
PE	0.084482
PU	0.505752
Sk-Ch	

TABLE IX. PATH COEFFICIENT AND T-VALUE

	Path Coefficient	Standard Error	t-value	p-value	t-table, α=0.05, df=185	Decision
Attitude -> Intention	0.633	0.137	4.628	0.000	1.653	**Supported**
Enjoyment -> Attitude	0.239	0.099	2.424	0.008		**Supported**
Intention -> Intensity	0.652	0.088	7.431	0.000		**Supported**
PC -> Attitude	0.066	0.111	0.594	0.277		**Not Support**
PC -> Intention	0.227	0.127	1.792	0.037		**Supported**
PE -> Attitude	0.343	0.137	2.509	0.006		**Supported**
PE -> PU	0.654	0.060	10.942	0.000		**Supported**
PU -> Attitude	0.232	0.109	2.118	0.018		**Supported**
Sk-Ch -> Enjoyment	0.660	0.078	8.429	0.000		**Supported**
Sk-Ch -> PC	0.470	0.081	5.801	0.000		**Supported**
Sk-Ch -> PE	0.291	0.105	2.756	0.003		**Supported**
Sk-Ch -> PU	0.147	0.082	1.794	0.037		**Supported**

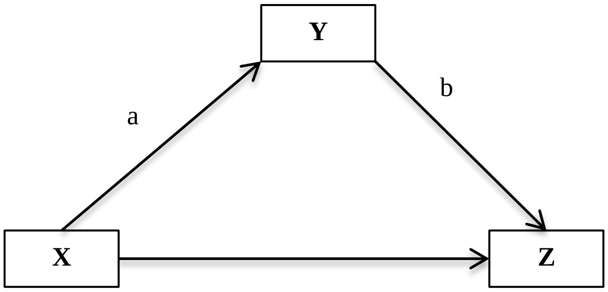


Fig. 4. Indirect effect.

Based on the data in Table IX, the value of standard normal distribution with Sobel test was obtained as shown in Table X.

By using 5% significance, the relation of indirect effect PE→PU→AT was significant, while indirect effect PC→AT→IT was not significant. This result was in line with values of path coefficient and t-value from relation PC→AT that was not significant. The results of testing the relationship between the learning behavior variable and the psychological experience variable are then used as the basis that the multicriteria attribute related to actual usage can be used to measure or predict psychological conditions in learning through e-learning.

TABLE X. PATH COEFFICIENT AND T-VALUE

Path	z value	p-value, one tailed	Decision
PE->PU->AT	2.080	0.019	Significant
PC->AT->IT	0.589	0.278	Not significant

VI. Conclusion

Investigation of relations among learning behavior as a multicriteria attribute, TAM, and flow experience as a representation of psychological experience was presented in this study. The main contribution of this research is to provide a way on how to measure the psychological experience of e-learning naturally. Learning behavior when interacting with e-learning becomes the main variable in determining attribute multicriteria. All variables illustrated in structure model showed positive relations, thus all causal relations proposed in the hypotheses were accepted. This result also provide indication that learning intensity as a representation of learning behavior could be used as a reference to identify user's behavior during interaction with e-learning. If this is combined with skill-challenge balance as an antecedent flow experience, it could create engagement between user and e-learning system. However, these parameters have not been tested in real cases. This research is still limited to the relationship between variables that can be used as attributes to measure or predict the psychological learning experience. In the future research, this multicriteria attribute can be used as parameters for prediction and classification of flow experience. It can use several prediction methods such as machine learning, rough set, or rough-regression.

Acknowledgment

This research is funded by the Grant of Fundamental Research Scheme from the Ministry of Research, Technology, and Higher Education, Republic of Indonesia, 2020.

References

[1] Manzoor, H.-L. Truong, and S. Dustdar, "Quality of Context: models and applications for context-aware systems in pervasive environments," Knowl. Eng. Rev., vol. 29, no. 2, pp. 154–170, Mar. 2014.

[2] G. Zhang and Z. Cheng, "A WWW-based learner's learning motivation detecting system," 2003.

[3] C. Whitson and J. Consoli, "Flow Theory and Student Engagement," J. Cross-Disciplinary Perspect. Educ., vol. 2, no. 1, pp. 40–49, 2009.

[4] J. Nakamura and M. Csikszentmihalyi, "The concept of flow," The handbook of positive psychology. pp. 89–105, 2002.

[5] D. J. Shernoff and E. Rowe, "Measuring Flow in Educational Games and Gamified Learning Environments Increasing Engagement in Learning through Serious Educational Video Games Theoretical Foundation : Flow Experiences and Their Relationship to Learning," pp. 2276–2281, 2012.

[6] M. Csikszentmihalyi, Flow: The Psychology of Optimal Experience. New York, NY: Harper and Row, 1990.

[7] P. Pu and L. Chen, "A user-centric evaluation framework of recommender systems," CEUR Workshop Proc., vol. 612, pp. 14–21, 2010.

[8] D. S. S. Sahid, L. E. Nugroho, and P. I. Santosa, "Modeling the Flow Experience for Personalized Context Aware E-learning," in Proceedings of The 8th 2016 International Conference on Information Technology and Electrical Engineering, 2016, pp. 236–241.

[9] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," MIS Q., vol. 13, no. 319–339, 1989.

[10] D. A. Adams, R. R. Nelson, and P. A. Todd, "Perceived Usefulness, Ease of Use, and Usage of Information Technology," MIS Q., vol. 16, no. 2, pp. 227–247, 1992.

[11] V. Venkatesh and F. D. Davis, "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies," Manage. Sci., vol. 46, no. 2, pp. 186–204, 2000.

[12] W. W. Chin and P. A. Todd, "On the Use, Usefulness, and Ease of Use of Structural Equation Modeling in MIS Research: A Note of Caution," MIS Q., vol. 19, no. 2, pp. 237–246, 1995.

[13] I. Ajzen, Attitudes, Personality and Behavior, 2nd ed. Open University Press, 2005.

[14] R. M. Gagne, L. J. Briggs, and W. W. Wager, Principles of Instructional Design. Orlando: Ted Buchholz, 1992.

[15] A. H. Eagly and S. Chaiken, The psychology of attitudes. Fort Worth, Tex. Harcourt Brace Jovanovich College, 1993.

[16] I. Ajzen and T. J. Madden, "Prediction of Goal-Directed Behavior: Attitudes, Intentions, and Perceived Behavioral Control," J. Exp. Soc. Psychol., vol. 22, pp. 453–474, 1986.

[17] N. S. Endler, R. L. Speer, J. M. Johnson, and G. L. Flett, "Controllability, coping, efficacy, and distress," Eur. J. Pers., vol. 14, no. 3, pp. 245–264, 2000.

[18] E. Skinner and T. Greene, "Perceived Control: Engagement, Coping, and Development," in 21st Century Education: A Reference Handbook, SAGE Publications Ltd, 2007, pp. 91–106.

[19] K. Bodey and D. Grace, "Examining Self-Monitoring, Perceived Control, Self-Efficacy and Machiavellianism in the Context of Complaint Behaviour," Mark. Accountabilities Responsib., 2004.

[20] M. Cocea and S. Weibelzahl, Log file analysis for disengagement detection in e-Learning environments, vol. 19, no. 4. 2009.

[21] G. B. Moneta, "On the Measurement and Conceptualization of Flow."

[22] L. Liao, "A Flow Theory Perspective on Learner Motivation and Behavior in Distance Education," Distance Educ., vol. 27, no. 1, pp. 45–62, 2006.

[23] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "School Engagement: Potential of the Concept, State of the Evidence," Rev. Educ. Res., vol. 74, no. 1, pp. 59–109, 2004.

[24] H. M. Marks, "Student Engagement in Instructional Activity: Patterns in the Elementary, Middle, and High School Years," Am. Educ. Res. J., vol. 37, no. 1, pp. 153–184, 2000.

[25] F. M. Newmann and G. G. Wehlage, "Five Standards of Authentic Instruction," Educ. Leadersh., vol. 50, no. 7, pp. 8–12, 1993.

[26] D. J. Shernoff, M. Csikszentmihalyi, B. Schneider, and E. S. Shernoff, "Student Engagement in High School Classrooms from the Perspective of Flow Theory," Sch. Psychol. Q., vol. 18, no. 2, pp. 158–176, 2003.

[27] M. C. Lee, "Explaining and predicting users' continuance intention toward e-learning: An extension of the expectation-confirmation model," Comput. Educ., vol. 54, no. 2, pp. 506–516, 2010.

[28] D. S. S. Sahid, R. Efendi, E.H. Putra and M. Wahyudi, "Categorizing Attributes in Identifying Learning Style Using Rough Set Theory," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 11, no. 1, 2020, pp. 292–298.

Determinants of Privacy Protection Behavior in Social Networking Sites

Siti Norlyana Suhaimi[1], Nur Fadzilah Othman[2]*, Raihana Syahirah[3]
Syarulnaziah Anawar[4], Zakiah Ayop[5], Cik Feresa Mohd Foozy[6]
Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Malaysia[1]
Center for Advanced Computing Technology, Fakulti Teknologi Maklumat dan Komunikasi
Universiti Teknikal Malaysia Melaka, Malaysia[2, 3, 4, 5]
Faculty of Computer Science and Information Technology, Universiti Tun Hussien Onn, Malaysia[6]

***Abstract*—Social Networking Sites (SNSs) are an attractive online platform for social interaction and communication. Since SNSs are easily accessed by a large number of people, a large quantity of data is also stored in the SNSs. Consequently, concern regarding the exposure to privacy risk will emerge. In this case, users need privacy protection behavior to protect their privacy in SNSs. This paper aims to determine the motivational determinants of privacy protection behavior among high school students in protecting their data or personal information when using SNSs. To identify the determinants of privacy protection behavior, a questionnaire survey was administered on 200 high school students. This study proposed a conceptual model that offers an understanding of motivational determinants of privacy protection behavior in social networking sites. Results indicate that perceived anonymity is the most significant determinant in motivating privacy behavior followed by perceived intrusiveness, perceived severity, self-efficacy, perceived vulnerability, and response efficacy. The results of this study will shed some light on understanding the levels of privacy protection behavior in SNSs, and identify suitable interventions in motivating privacy protection behavior among high school students. Finally, with the combined theory of Protection Motivation Theory (PMT) and Hyperpersonal Communication Theory (HCT), this model provides the basis to direct future studies in the related field.**

Keywords*—Privacy; social networking sites; privacy; protection motivation theory; hyperpersonal communication theory

I. Introduction

As technology evolves, the user of the technology also increases. Nowadays, people use technology and SNSs as a platform to build social networks by communicating with friends, knowledge sharing, updating others on their activities and whereabouts, sharing photos, videos, archiving events, getting updates on activities by friends, sending messages privately, and posting public testimonials. SNSs offer an attractive way for social interaction and communication, thus encourages public users to use them. The rapid development of current technology in addition to the various attractive SNSs applications is driving the increase in the number of users. SNSs have become a major platform in carrying out various activities.

SNSs technology is capable of storing and sorting huge quantities of data and is easily accessed by a large number of people [1], which may unnecessarily expose them to various threats. The wide variety of features in the SNSs will influence people to expose their data privacy by sharing their personal information when utilizing SNSs. Consequently, concerns regarding the exposure to privacy risks emerge. Acts and behaviors show by users when utilizing SNSs will affect their lives either positively or negatively. This is because action towards privacy essentially relies on the behavior of the user itself. If users were not careful while utilizing the SNSs, it will have a detrimental effect on their lives instead of adding beneficial effects. While many users state that they stay informed about the risks in SNSs, this does not necessarily mean that they have the skills or motivation to behave securely [2].

Therefore, this study aims to identify the determinants of privacy protection behavior when utilizing SNSs among high school students in Malaysia. By knowing the determinants of privacy protection behavior, it will be able to provide knowledge that can protect users and empower them to assert their self-control with confidence through the implementation of strategies for privacy protection.

II. Literature Review

A. *Privacy Protection Definition*

Privacy is the right or power to monitor the distribution or release of details about a user or their actions. Privacy can be hard to protect these days due to the user's lack of knowledge of privacy protection. Privacy security protects the user details to avoid slipping into the hands of hackers, political agencies, and other organizations [3]. The concept of privacy protection varies from person to person [4] because each individual has specific privacy standards, and the degree of protection they need to believe. There are scientific and analytical cases where privacy protection can both enhance and subtract from the wellbeing of people and communities. Privacy protection shall keep user personal data from individuals who may attempt to misuse it. Minimizing user digital footprint makes it easier for people to take advantage of user data and users. If an attacker has good or bad intentions, the user's digital footprint will say a lot about them. An attacker will predict what users are doing every day, what users are doing in their leisure time, where users are working, moreover user's social activities, and all of their personalities [5]. Thus, protecting privacy involves data protection against unwanted access [6].

*Corresponding Author

B. Privacy Issues of Social Networking Sites (SNSs)

Once an individual had involvement in SNSs, they will contribute to various privacy issues [7]. There are endless examples of data being collected without the consent of a person, identities being generated based on SNSs usage, accounts being hacked, etc. According to [8], the growing number of cases of fraud, identity theft, cyberbullying, cyberstalking, and others were seen as a crucial enabler for improving users' trust in using SNSs.

Instead of the personal information that usually got stolen from the SNSs user, the user always forgot that they may reveal too much of their private information by themselves by sharing photos or videos on the SNSs. In the context of Malaysia, there is nothing that the legal system in Malaysia can do to secure users when those kinds of data got stolen [9]. A survey by the MCA's Civil Protection and Grievances Office shows that the loss of RM4.5 million had been reported due to identity fraud from 2014 to 2017 [10].

Malaysian Communication and Multimedia Commission (MCMC) has advised SNSs users not to treat SNSs as a personal diary [11]. Through the monitoring of MCMC, it is found that users often share personal information, photos, and locations that can attract persons with malicious intent. According to statistics by [12] in Incident Statistic Report 2019, there were a total of 10,772 incidents reported in 2019, and fraud cases are the highest. SNSs are one of the mediums used by scammers to find the victims of their fraud activities. Personal information displayed on SNSs to some extent becomes a source of initial information in locating potential victims. Corresponds with the study by [13], it argues that the SNSs should be a place where the user could have fun to share about their life with everyone, but still secure and confidential. Nowadays it seems to be a minefield for an attacker to hack, steal private information, data monitoring, and networking exploitation. Thus, SNSs today is being more anti-social and intrusive of privacy.

III. Theoretical Framework

This section provides the theoretical framework explaining the concept and the determinants of privacy protection behavior in SNSs. The proposed theoretical framework is based on the Protection Motivation Theory and Hyperpersonal Communication Theory.

A. Protection Motivation Theory

The Protection Motivation Theory (PMT) offers analytical insight to explain the attraction of apprehension in individuals and the difference in actions against certain situations or environments [14]. The development of PMT was to understand the reason for users who are concerned about protecting themselves from potential risks and helping to enhance the knowledge of determinants that make a user conduct relevant, prescribed behavior and ensuring the safety of privacy. A rising number of studies have stated the PMT's importance to understand responses of the user to privacy threats on SNSs [15]. Prior research has already shown the significance of the perceptions and behavior regarding privacy. A user who is more engaged in privacy protection behavior is a user who cares and cautious about information privacy and attaches high priority to their privacy protection [14]. Furthermore, the more the user face the privacy issues, the higher the protective behavior goes [16].

PMT consists of five factors: Perceived Vulnerability (PV), Perceived Severity (PS), Self-Efficacy (SE), Response Efficacy (RE), and Reward (R). The PMT states that the motivation of users to protect themselves against particular threats is based on two matters which are a threat appraisal and a coping appraisal. For threat appraisal, it measures the perceived severity of the threats and the perceived vulnerability to those threats. For coping appraisal, it measures self-efficacy and response efficiency. Both of the appraisals affect the behavior of the user to protect their privacy from a threat. Those appraisals have connections when both of them are perceived as high, where users will have the encouragement and motivation to have protection behavior to protect them from the threats and change their behavior.

1) Perceived Severity (PS): Perceived severity is best described as the perceived seriousness of threatening outcomes. User changes their behavior according to the perceived severity of the consequence and thus reduce the risk of threats. Perceived severity generally refers to the user’s assumption that a threatening occurrence arises from a conclusion of severity significance [17]. Additionally, [18] stated that perceived severity can enhance a user’s willingness to participate in the behavior of lessening the threat. Simply put, a greater level of perceived severity intensity will force internet users to take protective measures in SNSs [19]. Although a significant effect of perceived severity on intention and behavior has been found by several researchers, there is also evidence that shows that the severity is not significantly related to intention [20].

2) Perceived Vulnerability (PV): Perceived vulnerability is the decision of a user as to the probability of a threat. Alternatively, perceived vulnerability describes when the user did experiences the negative impact in SNSs, it will make the user motivated and implement protection behavior [21]. Several studies support the notion of perceived vulnerability that has a beneficial impact on defense actions by users. Perceived vulnerability has been found to increase students' intent to perform threat avoidance behaviors [22]. Furthermore, if people find themselves more vulnerable to an adverse danger, they take defensive measures to mitigate the danger. Additionally, when individuals perceive themselves to be vulnerable to privacy risks, they seem to be more worried about their personal information [23], which causes an emotional reaction and anxiety, which in effect raises the desire for defense [24].

3) Self-efficacy (SE): Self-efficacy can be defined as the degree to which users believe they should implement the recommended behavior. Users should have the courage to resolve difficulties that prohibit them from taking a specific action. Studies found that in the sense of privacy environments, users must have specific technological skills. This is linked to one's desire to change one's unsafe or inefficient behavior. In his study, [25] argues that one can also strengthen one's actions toward more effective data protection initiatives with self-efficacy. When exchanging information, the user with better trust in their abilities to handle their privacy details may have fewer privacy issues [26]. Furthermore, [27] described self-efficacy as a core determinant of privacy protection or threat avoidance actions and a major factor in enhancing the effectiveness of the protection. This research attempts to identify the function of users' self-efficacy in the implementation of privacy protection behavior. As [28] stated that actual behavior is the main factor that affects self-efficacy. Hence, the role of the user's behavior as a basis of perceptions regarding self-efficacy has been largely established and confirmed.

4) Response Efficacy (RE): Response efficacy measures how effective the adoption of response in mitigating the threat. A study found that response efficacy is a major predictive activity that decides whether to introduce security measures on their networks or not, increases efforts to use anti-spyware as a protective tool, and predict backup of data on private computers [29]. Besides, [27] said that response efficacy is expected to take on a major role in reducing SNSa threats because the privacy of information is considered a question of ambiguity. Users that perceive improved privacy from a personalization program have less system-specific privacy concerns and are more likely to use it as a privacy tool [30]. Thus, the study concludes that having a good response efficacy would enable lower data loss.

5) Rewards (R): A study by [31] found that users reported getting a lot of credit on SNSs when a lot of personal information being posted. From previous research, rewards are described as receiving attention and response from posting on SNSs through like comments and "likes". However, implementing restrictive privacy settings can be a barrier to achieving certain rewards in SNSs [32]. Furthermore, users in SNSs trade their privacy information for gaining rewards in SNSs and obtain benefits such as popularity and enjoyment when disclosing personal information [33]. Once users experience the benefits of SNSs, they will choose to share their personal information to obtain these benefits [22]. Because those might give a negative impact on users as they desire to gain those rewards [34], users need to maximize the rewards derived from SNSs interactions with their contacts with privacy protection behavior within SNSs [35].

B. Hyperpersonal Communication Theory

Hyperpersonal Communication Theory (HCT) provides a model for understanding how users perceive emotional intimacy in computer-mediated communication (CMC)[36]. HCT provides more manageable online interaction, as information senders can cleverly select what and how to reveal [37]. Moreover, according to the concept of HCT, users had balanced their desires that sometimes competing for privacy with their willingness to be open in the SNSs environment and communicate with others. Previous research stated that the level and period of online messages could compensate and resulting in hyperpersonal communication or relationships that exceed face to face in terms of their emotional connection [36]. Hence, HCT identified that users of CMC are best able to discover their goodness by selecting the medium that fits their unique social needs perfectly [38].

1) Perceived anonymity of self: The perceived anonymity of self can be defined as the extent to which a sender perceives the source of the message as anonymous and undefined [38]. For instance, any picture, video, or post by a blogger or user on SNSs, could expose information about their virtual identity. Some identity information can be identified by at least their real name or their picture, while other things like blurred picture and nickname may only provide limited information about the user. At some point, the world of SNSs reveals users' real social identities and leads to a healthy exchange of content, whereas perceived anonymity of self-decreases the ability of users to share information [39]. The researcher has also found that a higher degree of perceived anonymity on SNSs means less need to reveal self [40]. Another study stated that perceived anonymity of self-gave a lot kind of good result, but somehow there is a lot of kind of user in which there must be some users give the negative effect of perceived anonymity of self [41]. As supported by some studies, the positive effect of anonymity of self on SNSs is the successful method to protect private information and create a private identity [42].

2) Perceived anonymity of others: The perceived anonymity of others relates to the absence of identification information and details about the other user on SNSs [43]. If other users are known as anonymous, recognizing who they are or keeping them to account for their acts and personal views is unlikely and difficult for the individual. The other user may be more likely to post information and comments using an anonymous online identity [44] and it makes individual difficult to know about the other user's identity as if the user is bad or not. The consequences of anonymity of others are mirrored in the results of the dark side of the Internet's personalization, fraud, and fake information [45], so individuals could probably fall into trap of scam. The anonymity of others can also present as a negative role in the exchange of information in internet-based interpersonal communication [46], as it always happens that the users fake their identity to make themselves feel better to communicate in SNSs. The view is supported by Twitter research that states that a lot of anonymous users have shared and create bad content by tweeting compare to non-anonymous users [47].

3) Perceived intrusiveness: Another serious recent problem in SNSs studies was perceived intrusiveness [48]. Perceived intrusiveness refers to how often people experience an unwanted violation of their environment [49]. In other words, high willingness in experiencing threats can lead to more sensitivity and less perceived intrusiveness [50]. Intrusiveness is a psychological theory that endorses the concept of creating an imbalance between the independence of two parties and the self-rule to protect personal identity on SNSs [51]. For instance, users could feel intrusive by spam from other users, slander, sexual harassment, cyberbully, advertisement, and many other intrusive things. This will create higher levels of perceived intrusiveness and therefore more negative emotions for people who are particularly anxious with privacy when presenting themselves than seeing others [52].

In order to better understand the reason for students to adopt privacy protection behavior in SNSs, there is a need to determine the factors of their adoption. The proposed framework shown in this study was constructed based on PMT for perceived severity, perceived vulnerability, self-efficacy, response efficacy and rewards. Besides that, three variables from HCT including perceived anonymity of self, perceived anonymity of others and perceived intrusiveness is also included as determinants of privacy protection behavior.

This study combines PMT and HCT because PMT is a basic theory that is often used in research related to privacy protection behavior while HCT is the theory that offers an approach to understand how users experience relational intimacy in computer mediated communication (CMC) [36]. The combination of these two theories with the addition of variables from HCT can add value to existing studies as it focuses more on computer-mediated communication.

C. Privacy Protection Behavior

Privacy protection refers to the management of personal information disclosure while deflecting unwanted intrusions [53]. Protecting privacy behavior in this modern era is a must since it makes users concern about how the data shared is being stored or collected, also how the data being exposed to others when the user shared it. SNSs users may protect their privacy by not exposing too much about themselves, learn how to limit the privacy information that they share and exchange also by taking security steps for privacy [54]. When individuals feel betrayed, sense of unfairness, inequality, and emotional distress, they would then adopt various privacy protection to protect their privacy [55]. Hence, there is a need to find out more about how users use privacy protection, so the outcomes of their actions can be determined.

Privacy protection refers to an action that individuals perform to keep their information safe and been categorized into two categories namely: i) approach strategies and ii) avoidance strategies [56]–[58]. Approach strategies refer to confrontation strategies that encompass problem-solving and seeking social support while avoidance strategies are withholding and refusing to provide the information. Some approach strategies including fabricating personal information and seeking social support by asking for information and advice or reading the privacy statement. An example of avoidance strategies is removing or deleting offending people in SNSs, using privacy settings provided by SNSs, choose and control who can see their profiles and their posts, and with whom they can share their personal information. The use of such privacy strategies is important so that they can make informed decisions about sharing their information in desired ways. Besides, the use of privacy setting can help SNSs users to reap the benefits from selectively sharing content on SNSs, while at the same time minimizing the potential damage and harm to their reputation and relationships that may result from unintentional disclosures [59].

This paper proposes the theoretical framework as shown in Fig. 1. Essentially, this study re-examines the constructs as privacy protection factors concerning information sharing in SNSs setting. Based on the previous discussions, the following hypotheses are proposed:

H1: Perceived severity positively influenced privacy protection behavior in SNSs.

H2: Perceived vulnerability positively influenced privacy protection behavior in SNSs.

H3: Self-efficacy positively influenced privacy protection behavior in SNSs.

H4: Response efficacy positively influenced privacy protection behavior in SNSs.

H5: Rewards positively influenced privacy protection behavior in SNSs.

H6: Perceived anonymity of self positively influenced privacy protection behavior in SNSs.

H7: Perceived anonymity of others positively influenced privacy protection behavior in SNSs.

H8: Perceived intrusiveness positively influenced privacy protection behavior in SNSs.

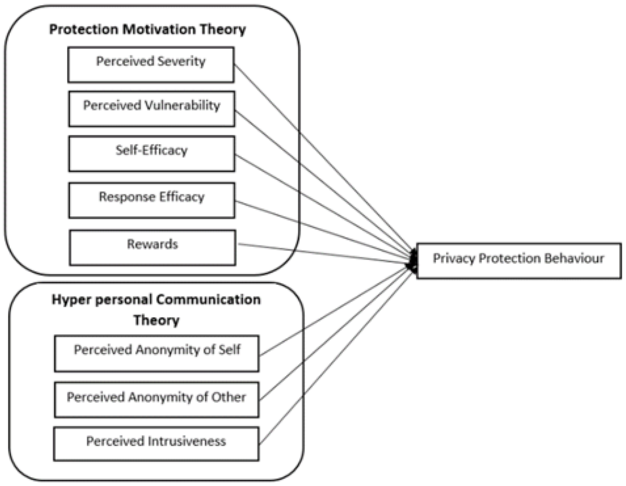

Protection Motivation Theory
Perceived Severity
Perceived Vulnerability
Self-Efficacy
Response Efficacy
Rewards
Hyper personal Communication Theory
Perceived Anonymity of Self
Perceived Anonymity of Other
Perceived Intrusiveness
Privacy Protection Behaviour

Fig. 1. Proposed Framework.

IV. Methodology

This section describes the details of the quantitative methodology adopted in the study.

A. Research Instrument Design

This study will use the survey questionnaire as the research instrument design. To detect any possible issues, the draft instruments were proposed to select experts in the field related. The aim is to remove any confusing or ambiguous words from the questionnaire, also help to improve questionnaire quality and reliability. Additionally, the questionnaire was designed to be easily followed and respond to reduce sampling errors and to have a high number of respondents that voluntarily answer the questionnaire.

To develop a questionnaire, it is necessary to identify the variables. Questions are formulated depending on the appropriateness of the variables. Every defined variable consists of approximately ten items. To calculate the number of elements, the Likert scale is used as the method. The function of the Likert scale is to allow individuals to show an agreement level on a particular statement with a five-point scale. The scale used for topical value variable is 1 being "Strongly Disagree", 2 being "Disagree," 3 being "Neutral," 4 being "Agree" and 5 as "Strongly Agree". The questionnaire was divided into three sections, which are Section A for Demographic Information, Section B for Privacy Protection Behavior, and Section C for determinants that motivate Privacy Protection Behavior.

B. Content Validity

The aim of validating the content is to identify the items which represent the variable in the generated questionnaire. To conduct validation, selected experts will be given the form of content validation as annexed in the appendix. Three information security experts were involved in the testing of this study. The experts were selected based on their computer security expertise, with at least five to twenty years' experience in the field related to the study. The content validation of the research instrument was statistically analyzed using a Content Validation Index (CVI). CVI was stated as the higher common approach apply to quantitatively measure the validity of the content. In this validation case, it has two types of CVI, which is Item-CVI (I-CVI) and Scale-CVI (S-CVI). Simply described I-CVI is the total agreed rated by experts while S-CVI will be the average of total I-CVI. Thus, the item rated less than 1.00 in I-CVI, the item will be excluded from the set of questionnaires and the S-CVI must be higher than 0.90 for the variable of the item validated.

The result in Table I shows that two items in PPB, PS, SE, and PI were removed, three items in PV, one item in RE and PAO, four items in R, and no items were removed in PAOS and IPC. The total number of items removed was 17 and 54 items were retained in the questionnaire.

C. Data Collection

For the data collection, the target respondent as a sample of the population is high school students in Malaysia who are the users of SNSs. The questionnaire on the survey instrument is circulated online, using the Google Form. Using shared links through an online platform and social media, the questionnaire will be distributed through SNSs such as Whatsapp, Instagram, and Telegram. The data collection is completed in two weeks. All answers are gathered and recorded before analyzing the results.

The sampling method that is used for data collection is non-proportional quota sampling. This study target 200 respondents from different schools with computer courses, the age range are 16 to 19 years old. To take part in this study a total of 230 respondents must be chosen. From a total of 230 respondents, the first 30 respondents were involved in pilot tests and will be excluded from the main study. The remaining 200 respondents will go to the actual survey. To maximize accuracy, students who registered on at least one SNSs were determined to select the students to be involved in this study. Those who had no SNSs accounts would not be included in the study sample.

D. Reliability

Reliability was achieved by analyzing the items from the pilot test result and obtain the Cronbach's Alpha value. The value of Cronbach's Alpha was determined based on the pilot test result. The general Cronbach's Alpha value should be higher than 0.70. The higher the score the greater the reliability of the scale produced. For Section B, variable privacy protection behavior shows a result of 0.742 which means the internal consistency is respectable. Besides that, Section C contains 9 variables which show the coefficient alpha result as follows: perceived severity (0.814), perceived vulnerability (0.800), self-efficacy (0.810), response efficacy (0.891), perceived anonymity of self (0.857), and perceived intrusiveness (0.847). The results with an alpha value above 0.8 mean the internal consistency is very good. The results with an alpha value above 0.90 mean the internal consistency is excellent: rewards (0.917), perceived anonymity of others (0.925), and information privacy concern (0.934). Hence, overall from this pilot study, the questionnaire was tested as suitable for actual study in this research. No variable needs to be removed for the actual study. Table II shows the results of Cronbach's Alpha value.

TABLE I. CVI RESULTS

Variable	S-CVI (Before)	S-CVI (After)
PPB	0.906	1.00
PS	0.906	1.00
PV	0.810	1.00
SE	0.906	1.00
RE	0.945	1.00
R	0.624	1.00
PAOS	1.00	1.00
PAO	0.953	1.00
PI	0.890	1.00

TABLE II. RESULTS OF CRONBACH'S ALPHA VALUE

Section	Sub-Construct	Alpha
B	Privacy Protection Behavior	0.742
C	Perceived Severity	0.814
	Perceived Vulnerability	0.800
	Self-Efficacy	0.810
	Response Efficacy	0.891
	Rewards	0.917
	Perceived Anonymity of Self	0.857
	Perceived Anonymity of Others	0.925
	Perceived Intrusiveness	0.847

V. Results

In this section, the result obtained from the survey will be analyzed through a few analyses.

A. Factor Analysis

In this study, Principal Component Analysis is used to perform construct validity. During the analysis, the items with low load factor values will be removed as it was considered as problematic. In this study, it had been set a higher cut-off value of 0.6 for loading factors [60]. The items removed were 8 items which are PAOS1, PAOS2, PAOS3, PAOS4, PAO1, PAO2, PAO3, and PI3 with factor loadings of less than 0.6. Factors that contain less than 3 items are counted as useless and weak, so it must be eliminated [61]. In this case, there is no factor in less than 3 items, hence, no factor will be removed. The final analysis shows that 41 items were retained. Also, two factors fall under the same number of the component which is number 4, PAOS and PAO. Both of the factors were a different factor but were combined under one factor. PAOS is the individual being anonymous in SNSs while PAO is another user being anonymous in SNSs. Assuming that, high school student believes that it is no difference between both of the factor, either themselves of other user being anonymous in SNSs might help in protecting the privacy. By being anonymous, they could implement privacy protection behavior. Therefore, in this case, both of them were combined under one factor namely PA (Perceived Anonymity). Table III shows the results of the factor analysis.

TABLE III. Results of Factor Analysis

Item	1	2	3	4	5	6	7
PS1				.715			
PS2				.787			
PS3				.688			
PS4				.678			
PS5				.723			
PV1					.690		
PV2					.786		
PV3					.871		
PV4					.851		
SE1		.692					
SE2		.798					
SE3		.720					
SE4		.723					
SE5		.749					
RE1	.737						
RE2	.771						
RE3	.658						
RE4	.808						
RE5	.824						
R1						.729	
R2						.894	
R3						.861	
R4						.659	
PAOS5			.714				
PAOS6			.660				
PAOS7			.662				
PAO4			.722				
PAO5			.578				
PAO6			.660				
PI1							.697
PI2							.695
PI4							.675

B. Regression Analysis

Multiple regression analysis is conducted to estimate the relationship between some of the independent variables towards a dependent variable. Multiple regressions were run in this study and the results are shown in Table IV. The results of multiple regression analysis showed that six independent variables, i.e. perceived severity, perceived vulnerability, self-efficacy, response efficacy, perceived anonymity, and perceived intrusiveness significantly influenced the dependent variable which is privacy protection behavior. However, H5 was not supported in this study. Based on the result obtained, the model that predicts the motivational determinants of privacy protection behavior were identified. The β value shows the strength of the relationship, the higher the β value the stronger the relationship. R square for this regression model was 0.331, which indicated 33.1% of the variance in privacy protection behavior is explained by perceived severity, perceived vulnerability, self-efficacy, response efficacy, perceived anonymity, and perceived intrusiveness. Further examination on the standardized beta coefficient revealed that the most dominant factor that affects the respondents' privacy protection behavior was perceived anonymity (β = .378), followed by perceived intrusiveness (β = .277), perceived severity (β = .264), self-efficacy (β = .227), perceived vulnerability (β = .210) and response efficacy (β = .209).

TABLE IV. Results of Multiple Regression

Model	**Standardized Coefficients**	***t***	***Sig.***
	Beta (β)		
PS → PPB	.264	3.977	.000
PV → PPB	.210	3.360	.001
SE → PPB	.227	3.202	.002
RE → PPB	.209	3.229	.001
R → PPB	.095	1.499	.135
PA → PPB	.378	5.946	.000
PI → PPB	.277	4.352	.000
Notes: Overall Model F= 48.334; p<0.05; R^2 = 0.331; adjusted R^2 = 0.34			

VI. Discussion

From the factor analysis, a few items were removed because of the low load factor as it was considered problematic. Also, two variables were merged into one factor namely Perceived Anonymity as it measures the same thing which is Perceived Anonymity of Self and Perceived Anonymity of Other. From these results, we can assume that high school students believed that anonymity does not matter, whether, on their side or the side of others, does help to keep their personal information safe.

As for the result of regression analysis, the potential of motivational determinants was perceived severity, perceived vulnerability, self-efficacy, response efficacy, perceived anonymity, and perceived intrusiveness and it was determined as positively influenced with privacy protection behavior. The determinants do motivate high school students to implement privacy protection behavior. Perceived anonymity was found to be the highest determinant of privacy protection behavior. This result shows that the students highly agreed that even if they try to hide their identity, their privacy can still be disturbed. In SNSs, even if users try to conceal their identity using nicknames or images that do not show their identity, other users might be able to guess their identity based on their mutual friends. Besides, students are also motivated to adopt protective action when confronted with strangers or anonymous people on SNSs. This is perhaps when others decline to expose their identification, students find it difficult to get enough factual information to better understand others. Unknown individuals might have bad intentions. Hence, it motivates students to adopt privacy protection behavior on SNSs even if the identity is anonymous.

VII. CONCLUSION

The main purpose of this research is to study the potential motivational determinants of privacy protection behavior. A model for the privacy protection behavior in SNSs is proposed by defining the main elements and provide a comprehensive model that will motivate the high school students in implementing privacy protection behavior. The results of this study are expected to be used to increase the level of privacy protection behavior among all users in SNSs and also to build up privacy guidelines. This study could motivate and influence the user to implement privacy protection behavior when utilizing SNSs.

Financial and time constraints limit the selection of the population in this study. However, the generalization for this study can be applied to the level of all students who have similar characteristics. The recommendation for future study in this field could overcome the limitation of this study such as the method to collect data by using quota sampling. This could be overcome by using a probability sampling method that could avoid bias selection on the population. The method of collecting the feedback from the respondent by using a questionnaire also could be improved in a future study to gain more variety of feedback such as interview, recording, observation, and others. A better and specific study to gain more understanding of privacy protection behavior can be obtained using qualitative analysis. Besides, this study could be improved by adding more motivational determinants towards privacy protection behavior. There could be more potential motivational determinants to exert more significant influence and impact on privacy protection behavior. Last but not least, expanding sample size and other ages with their background of education or various experience of the respondent can be extended to better generalize the analysis and potentially strengthen it among users of SNSs in Malaysia. By expanding sample size and age with educational background or experience, it may result in different intend, patterns, and behavior in SNSs. Hence, the future study can overcome all the limitations in this study to gain better and specific results of privacy protection behavior among SNSs users in Malaysia.

ACKNOWLEDGMENT

The authors would like to thank Center for Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM) for supporting the work done in this paper.

REFERENCES

[1] Kassim, P.N.J., 2008. The Development of e-Health in Malaysia: New Challenges to the Healthcare Industry.

[2] Zilles, M., 2011. Online Social Networks in Germany: Privacy Behaviour and Concern.

[3] S. Encrypt, “What is Privacy Protection? [Updated for 2020],” Search Encrypt Blog, Dec. 20, 2019.

[4] Van den Hoven, Jeroen, et al. “Privacy and Information Technology.” The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Summer 2020.

[5] M. Micheli, C. Lutz, and M. Büchi, “Digital Footprints: An Emerging Dimension of Digital Inequality,” Journal of Information Communication and Ethics in Society, vol. 16, pp. 242–251, Jun. 2018.

[6] R. Robinson, “Data Privacy vs. Data Protection,” Jan. 30, 2020.

[7] F. O., “Social Networking Privacy Concerns Impacting Businesses and Consumers,” Security Boulevard, N, 2019.

[8] Shanthi Kandiah, “Malaysia - The Privacy, Data Protection and Cybersecurity Law Review - Edition 6 - TLR - The Law Reviews,” Oct. 2019.

[9] Royce Tan and Sharmila Nair "M’sia sees biggest mobile data breach", NATION, Tuesday, 31 Oct 2017.

[10] Khairani Afifi Noordin, “News: Protect yourself from identity theft,” The Edge Markets, Nov. 21, 2017.

[11] Bernama, 2015. Laman Sosial Bukan Diari Peribadi. Malaysian Communication and Multimedia Commission. Available at: http://www.skmm.gov.my/Media/Press-Clippings/Laman-Sosial-Bukan-Diari-Peribadi-SKMM.aspx?lang=en-US.

[12] Malaysia Computer Emergency Report Team (MyCERT), 2020. Reported Incidents based on General Incident Classification Statistics 2019. Available at: https://www.mycert.org.my/portal/statistics-content?menu=b75e037d-6ee3-4d11-8169 66677d694932&id=0d39dd96-835b-44c7-b710-139e560f6ae0

[13] S. C. Boerman, S. Kruikemeier, and F. J. Zuiderveen Borgesius, “Exploring Motivations for Online Privacy Protection Behavior: Insights From Panel Data,” Communication Research, Oct. 2018.

[14] L. Baruh, E. Secinti, and Z. Cemalcilar, “Online Privacy Concerns and Privacy Management: A Meta-Analytical Review,” J Commun, vol. 67, no. 1, pp. 26–53, Feb. 2017.

[15] M. Büchi, N. Just, and M. Latzer, “Caring is not enough: the importance of Internet skills for online privacy protection,” Information, Communication & Society, vol. 20, no. 8, pp. 1261–1278, Aug. 2017.

[16] M. C. Green, “Social Network Sites and Well-Being: The Role of Social Connection,” Curr Dir Psychol Sci, vol. 27, no. 1, pp. 32–37, Feb. 2018.

[17] B. Palladino et al. and E. Menesini et al., “Perceived Severity of Cyberbullying: Differences and Similarities across Four Countries,” Frontiers in Psychology, vol. 8, p. 1524, Sep. 2017.

[18] K. Adhikari and R. K. Panda, “Users’ Information Privacy Concerns and Privacy Protection Behaviors in Social Networks,” Journal of Global Marketing, vol. 31, no. 2, pp. 96–110, Mar. 2018.

[19] T. Wang, T. D. Duong, and C. C. Chen, “Intention to disclose personal information via mobile applications: A privacy calculus perspective,” International Journal of Information Management, vol. 36, no. 4, pp. 531–542, Aug. 2016.

[20] B. Hanus and Y. Wu, “Impact of Users’ Security Awareness on Desktop Security Behavior: A Protection Motivation Theory Perspective,” Information Systems Management, vol. 33, Nov. 2015.

[21] S. Jain and S. Agrawal, “Perceived vulnerability of cyberbullying on social networking sites: effects of security measures, addiction and self-disclosure,” Indian Growth and Development Review, vol. ahead-of-print, Jun. 2020.

[22] M. Abdul Hameed and N. Arachchilage, “On the Impact of Perceived Vulnerability in the Adoption of Information Systems Security Innovations,” International Journal of Computer Network and Information Security, vol. 11, pp. 9–18, Apr. 2019.

[23] X. Zhang, S. Liu, X. Chen, L. Wang, B. Gao, and Q. Zhu, “Health information privacy concerns, antecedents, and information disclosure intention in online health communities,” Information & Management, vol. 55, no. 4, pp. 482–493, Jun. 2018.

[24] F. Mwagwabi, T. McGill, and M. Dixon, “Short-term and Long-term Effects of Fear Appeals in Improving Compliance with Password Guidelines,” Communications of the Association for Information Systems, pp. 147 – 182, Feb. 2018.

[25] M. Vatka, “Information Behavior And Data Security : Health Belief Model Perspective,” 2019.

[26] N. Arachchilage, “User-Centred Security Education: A Game Design to Thwart Phishing Attacks,” Nov. 2015.

[27] R. Fida, C. Tramontano, M. Paciello, Valerio Ghezzi, and C. Barbaranelli, “Understanding the Interplay Among Regulatory Self-Efficacy, Moral Disengagement, and Academic Cheating Behavior During Vocational Education: A Three-Wave Study,” Journal of Business Ethics, vol. 153, Dec. 2018.

[28] K. D. Martin, A. Borah, and R. W. Palmatier, “Data Privacy: Effects on Customer and Firm Performance,” Journal of Marketing, vol. 81, no. 1, pp. 36–58, Jan. 2017.

[29] H. Lee and A. Kobsa, “Understanding user privacy in Internet of Things environments,” in 2016 IEEE 3rd World Forum on Internet of Things (WF-IoT), Reston, VA, USA, Dec. 2016, pp. 407–412.

[30] T. Dienlin and M. J. Metzger, “An Extended Privacy Calculus Model for SNSs: Analyzing Self-Disclosure and Self-Withdrawal in a Representative U.S. Sample,” Journal of Computer-Mediated Communication, vol. 21, no. 5, pp. 368–383, 2016.

[31] D. Wang, “A study of the relationship between narcissism, extraversion, body-esteem, social comparison orientation and selfie-editing behavior on social networking sites,” Personality and Individual Differences, vol. 146, pp. 127–129, Apr. 2019.

[32] H.-T. Chen and W. Chen, “Couldn’t or Wouldn’t? The Influence of Privacy Concerns and Self-Efficacy in Privacy Management on Privacy Protection,” Cyberpsychology, behavior and social networking, vol. 18, pp. 13–9, Jan. 2015.

[33] N. Park and Y.-J. Kim, “The Impact of Social Networks and Privacy on Electronic Word-of-Mouth in Facebook: Exploring Gender Differences,” 2020.

[34] Craig R.Scott, “To Reveal or Not to Reveal: A Theoretical Model of Anonymous Communication,” Commun Theory, vol. 8, no. 4, pp. 381–407, Nov. 1998.

[35] Z. Liu, Q. Min, Q. Zhai, and R. Smyth, “Self-disclosure: A social exchange theory perspective,” Information & Management, vol. 53, no. 1, pp. 53–63, Jan. 2016.

[36] E. Sumner and A. Ramirez, “Social Information Processing Theory and Hyperpersonal Perspective,” 2017.

[37] J. Mou and D. Shin, “Effects of social popularity and time scarcity on online consumer behavior regarding smart healthcare products: An eye-tracking approach,” Computers in Human Behavior, vol. 78, pp. 74–89, Jan. 2018.

[38] T. L. Cigelske, “The Highest Form of Like: Snapchat, College Students and Hyperpersonal Communication,” undefined, 2018.

[39] X. Chen, M. Sun, D. Wu, and X. Y. Song, “Information-Sharing Behavior on WeChat Moments: The Role of Anonymity, Familiarity, and Intrinsic Motivation,” Front. Psychol., vol. 10, 2019.

[40] C. P. Barlett, D. A. Gentile, and C. Chew, “Predicting cyberbullying from anonymity.,” Psychology of Popular Media Culture, vol. 5, no. 2, pp. 171–180, Apr. 2016.

[41] X. Lin, M. Featherman, and S. Sarker, “Understanding factors affecting users’ social networking site continuance: A gender difference perspective,” Information & Management, vol. 54, no. 3, pp. 383–395, Apr. 2017.

[42] L. Levontin and E. Yom-Tov, “Negative Self-Disclosure on the Web: The Role of Guilt Relief,” Frontiers in Psychology, vol. 8, Jun. 2017.

[43] H. Nissenbaum, “The Meaning of Anonymity in an Information Age,” The Information Society, vol. 15, no. 2, pp. 141–144, May 1999.

[44] E. Jardine, “The Dark Web Dilemma: Tor, Anonymity and Online Policing,” Global Commission on Internet Governance Paper Series, p. 24, Sep. 2015.

[45] J. Fox and J. J. Moreland, “The dark side of social networking sites: An exploration of the relational and psychological stressors associated with Facebook use and affordances,” Computers in Human Behavior, vol. 45, pp. 168–176, 2015.

[46] J. D. Morris, Y. Choi, and I. Ju, “Are Social Marketing and Advertising Communications (SMACs) Meaningful?: A Survey of Facebook User Emotional Responses, Source Credibility, Personal Relevance, and Perceived Intrusiveness,” Journal of Current Issues & Research in Advertising, vol. 37, no. 2, pp. 165–182, Jul. 2016.

[47] J. Burgoon, R. Parrott, B. Poire, D. Kelley, J. Walther, and D. Perry, “Maintaining and Restoring Privacy Through Communication in Different Types of Relationships,” Journal of Social and Personal Relationships - J SOC PERSON RELAT, vol. 6, pp. 131–158, May 1989.

[48] N. A. Dodoo and J. (Taylor) Wen, “Weakening the avoidance bug: The impact of personality traits in ad avoidance on social networking sites,” Journal of Marketing Communications, vol. 0, no. 0, pp. 1–24, Jan. 2020.

[49] V. M. Wottrich, E. A. van Reijmersdal, and E. G. Smit, “App Users Unwittingly in the Spotlight: A Model of Privacy Protection in Mobile Apps,” Journal of Consumer Affairs, vol. 53, no. 3, pp. 1056–1083, 2019.

[50] Y. Feng and Q. Xie, “Privacy Concerns, Perceived Intrusiveness, and Privacy Controls: An Analysis of Virtual Try-On Apps,” Journal of Interactive Advertising, pp. 1–41, Sep. 2018.

[51] C. Goodwin, “A Conceptualization of Motives to Seek Privacy for Nondeviant Consumption,” J. Consum. Psychol., vol. 1, no. 3, pp. 261–284, 1992.

[52] M. Qi and D. Edgar-Nevill, “Social networking searching and privacy issues,” Inf. Secur. Tech. Rep., vol. 16, no. 2, pp. 74–78, 2011, doi: 10.1016/j.istr.2011.09.005.

[53] B. C. F. Choi, Z. Jiang, B. Ramesh, and Y. Dong, “Privacy tradeoff and social application usage,” *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, vol. 2015-March, pp. 304–313, 2015, doi: 10.1109/HICSS.2015.44.

[54] Y. Feng and W. Xie, “Teens’ concern for privacy when using social networking sites: An analysis of socialization agents and relationships with privacy-protecting behaviors,” Comput. Human Behav., vol. 33, no. July, pp. 153–162, 2014, doi: 10.1016/j.chb.2014.01.009.

[55] B. C. F. Choi, Z. Jiang, B. Ramesh, and Y. Dong, “Privacy tradeoff and social application usage,” Proc. Annu. Hawaii Int. Conf. Syst. Sci., vol. 2015-March, pp. 304–313, 2015, doi: 10.1109/HICSS.2015.44.

[56] E. G. Smit, G. Van Noort, and H. a. M. Voorveld, “Understanding online behavioral advertising: User knowledge, privacy concerns and online coping behavior in Europe,” Comput. Human Behav., vol. 32, pp. 15–22, Mar. 2014, doi: 10.1016/j.chb.2013.11.008.

[57] E. Litt, “Understanding social network site users’ privacy tool use,” Comput. Human Behav., vol. 29, no. 4, pp. 1649–1656, 2013, doi: 10.1016/j.chb.2013.01.049.

[58] Tu, Z., Turel, O., Yuan, Y. & Archer, N., 2015. Learning to cope with information security risks regarding mobile device loss or theft: An empirical examination. Information and Management, 52(4), pp.506–517.

[59] F. Banhawi, N. Mohamad Ali, and H. Judi, “User engagement attributes and levels in Facebook,” Journal of Theoretical and Applied Information Technology, vol. 15, Jan. 2012.

[60] R. Maskey, J. Fei, and H.-O. Nguyen, “Use of Exploratory Factor Analysis in Maritime Research,” The Asian Journal of Shipping and Logistics, vol. 34, no. 2, Art. no. 2, 2018.

[61] T. W. MacFarland and J. M. Yates, Introduction to Nonparametric Statistics for the Biological Sciences Using R. Springer International Publishing, 2016.

A Hybird Framework based on Autoencoder and Deep Neural Networks for Fashion Image Classification

Aziz Alotaibi
College of Computers and Information Technology, Computer Science Department
Taif University, Taif 21974, Saudi Arabia

***Abstract*—Deep learning has played a huge role in computer vision fields due to its ability to extract underlying and complex features of input images. Deep learning is applied to complex vision tasks to perform image recognition and classification. Recently, Apparel classification, is an application of computer vision, has been intensively explored and investigated. This paper proposes an effective framework, called DeepAutoDNN, based on deep learning algorithms for apparel classification. DeepAutoDNN framework combines a deep autoencoder with deep neural networks to extract the complex patterns and high-level features of fashion images in supervised manner. These features are utilized via categorical classifier to predict the given image to the right label. To evaluate the performance and investigate the efficiency of the proposed framework, several experiments have been conducted on the Fashion-MNIST dataset, which consists of 70000 images: 60000 and 10000 images for training and test, respectively. The results have shown that the proposed framework can achieve accuracy of 93.4%. In the future, this framework performance can be improved by utilizing generative adversarial networks and its variant.**

***Keywords*—Fashion detection; fashion classification; convolutional autoencoder; deep learning; insert**

I. Introduction

Conventional machine learning algorithms still used in the image processing and computer vision to perform data mining and feature extraction such as support vector machine (SVM). However, conventional machine learning algorithms have some limitations when dealing 1) with unstructured large-scale data and 2) the utilization of available advance computer resources with graphic processing unit (GPU) and tensor processing unit (TPU) [1]. With the great advancement in deep learning techniques exploring and solving the most complicated computer vision tasks, many computer vision applications have gained a significant attention due to the available resources and large amount of data. Therefore, these factors allow for the acceleration of models' training and for alleviating of the vanish gradient through creating a deep learning model such as ResNet50. In addition, deep learning algorithms directly process raw data (e.g., an RGB image) and obviate the need for preprocessing phase and domain experts. The aim of feature engineering is to learn useful representations in order to make a better decision. Deep learning algorithms is divide into two types: discriminative learning algorithms and generative learning algorithms. In this study, we focus only on discriminative learning algorithms such convolution neural networks (CNN), recurrent neural networks (RNN), Neural network (NN), and a few to name. In the past few years, deep learning have been intensively searched for visual data processing including image classification, image detection, semantic segmentation, and video processing [2] such apparel classification. More recently, with the great development in online shopping and e-commerce, visual fashion analysis and recognition has attracted many researches to utilize both computer vision and deep learning techniques. Apparel detection and classification is becoming one of the most used computer vision applications in online industry for some advantages: first, online shop recognition and recommendation of desirable fashion products [3-5]. Second, enhancing user experience [6, 7]. Third, improving online advertising [8, 9]. Finally, improving the performance of human detection and recognition in different scenarios [10, 11]. However, apparel fashion applications still encounter some difficulties to perform well such as; 1) the variation of the taken image size, 2) the light, 3) the angle, 4) new style, 5) undefined subcategories, 6) . Furthermore, due to the complex pattern of fashion apparel, various fashion properties, and the performance of previous existing deep learning algorithms, we believe that applying a good representation learning technique to extract useful features would enhance the performance of the fashion apparel detection and classification. Therefore, we propose a novel framework utilizing deep spatial autoencoder. Therefore, the proposed framework has proven to extract robust representations and effective in classifying fashion images with accuracy of 93.4%.

The main contribution of this study can be summarized as follows:

1) Propose a novel fashion classification framework based on deep learning techniques.

2) Utilize the deep autoencoder to reduce the dimensionality of the input fashion image and deep neural networks as classifier.

3) The proposed framework achieves an accuracy of 93.4% on Fashion MNIST which outperforms some of the existing deep learning algorithms such as CNN.

4) The proposed framework are evaluated through several experimental analyses.

The main contribution of this work is that it proposes a novel framework for fashion classification. The remainder of this paper is organized as follows: Section 2 discusses the previous work related to the proposed framework. In Section 3, the proposed framework including both deep autoencoder and deep neural networks are demonstrated and explained. Section 4 illustrates and discusses the evaluation performance. Finally, Section 5 concludes this work and introduces future works.

II. Related Work

In this section, a literature review is conducted and only covered the most recent related studies that mainly employee the machine learning and deep learning algorithms on Fashion MNIST dataset and its applications. Soe at el. [1] proposed a Hierarchical Convolutional Neural (H–CNN) for fashion image classification which emphasize on the knowledge embedded classifier outputting hierarchical information. The H-CNN is utilizing both VGG16 and VGG19 as base model which composed of five convolutional layers, max-pooling layers and fully connected layers for both feature extraction and classification. In addition, Duana et al. [12] utilized VGG-11 model which consists of convolution layers, maximum pooling layers and followed by batch normalization layers to classify Fashion images. In [13], authors presented a deep convolutional neural network (ConvNet) to classify the fashion images. In addition, author in [14] proposed a CNN model with Support Vector Machine for Fashion images classification. The author in [15] introduced various Hyper-Parameter Optimization (HPO) methods and regularization techniques using four deep convolution layers for recognizing images of fashion objects. Bhatnagar et al. [7] presented a three different deep convolution network with batch normalization and residual skip connection to acceleration the learning process. Their model reports enhanced accuracy of around 2% over the other deep learning systems. Authors [11, 16] proposed a deep Capsule Network, DeepCaps, which uses the concepts of skip connections and 3D convolutions. The skip connection allows a good gradient within a Capsule cell during the backpropagation optimization. Authors utilized the novel dynamic routing algorithm to assist the learning process of deep Capsule network. In addition, Deliege et al. [17] introduced a deep learning networks, called HitNet, with capsules embedded for data augmentation. HitNet uses a hybrid Hit-or-Miss layer to synthesis the representative images of a specific class by utilizing a reconstruction network. In [18], authors presented a fashion images classification system using single feature descriptor, histogram of oriented gradient (HOG), and multiclass support vector machine (SVM) for classification and detection of fashion images utilizing Fashion MNIST dataset. Shen [19] utilized the Long Short Term Memory (LSTM) to construct a model that can perform a classification task on Fashion MNIST dataset. Li et al. [20] proposed a personalized representation learning that is used for transforming shared feature extractors into personalized feature extractors. Their proposed study is based on two variants of Collaborative neural networks: Unconditional Collaborative Neural Networks (U-CoNN) and Conditional Convolutional Neural Networks (C-CoNN) for fashion image classification. In addition, Xiao et al. [21] introduced a multi-class object detection Coprocessor by combining Histogram of Oriented Gradient (HOG) feature and Local Binary Pattern (LBP) feature with a weighted Softmax classifier for image detection task such as Fashion MNIST.

III. Proposed Methodology

The aim of this proposed framework is to enhance the performance of the fashion classification. The proposed DeepAutoDNN framework based on deep learning techniques for fashion classification is illustrated in Fig. 1. The framework consists of three parts: First, deep autoencoder, second, deep neural networks and finally the classifier.

```
Algorithm 1
1. Input: training image set $\{x_i\}_1^n$
2. Deep Autoencoder computation
  A. Initialization of variable
  B. Compute encoder and decoder using Equation (1)(2)
  C. Minimize the construction error using Equation (3)
  D. Compute the latent representation $\{y_i\}_1^n$and update the weights
  E. Repeat step B, C, and D until convergence
  F. return $\{y_i\}_1^n$
3. Deep neural network
  A. input: the output of $\{y_i\}_1^n$
  B. train deep neural networks
  C. minimize the training error via backpropagation
4. softmax classifier
  A. predicted label y using weight in step 2 and 3
```

A. Deep Autoencoder

Autoencoder is an unsupervised representation learning algorithm based on artificial neural networks that composed of three typical layers; input layer, hidden layer and output layer [22]. The key goal of the autoencoder is to identify and disentangle the underlying hidden representation for a set of data. The process training of autoencoder consists of two types: encoder and decoder as shown in Fig. 2. The goal of the encoder is to map the input data to hidden representation, and the decoder is to reconstruct the original data from the hidden representation. The set of data is represented as $\{x_n\}_{n=1}^{N}$, where $X_n \in \mathbb{R}^m$, h_n donates the hidden representation, and $\hat{x}$ represents the reconstructed output. The encoder and decoder are calculated in equation 1 and 2 respectively as follows:

$$h_n = f\,(W_1 x_1 + b_1) \tag{1}$$

$$\hat{x} = g\,(W_2\, h_n + b_2) \tag{2}$$

Where f and g is the encoder and decoder function respectively, and W_1, W_2 donates the weight matrix of the encoder and decoder respectively.

The parameter sets of the autoencoder are optimized to minimize the reconstruction error:

$$(W_1, W_2) = \sum_{j\in\,\Omega_i} L(x, \hat{x}) \tag{3}$$

Where L donates the reconstruction error [23].

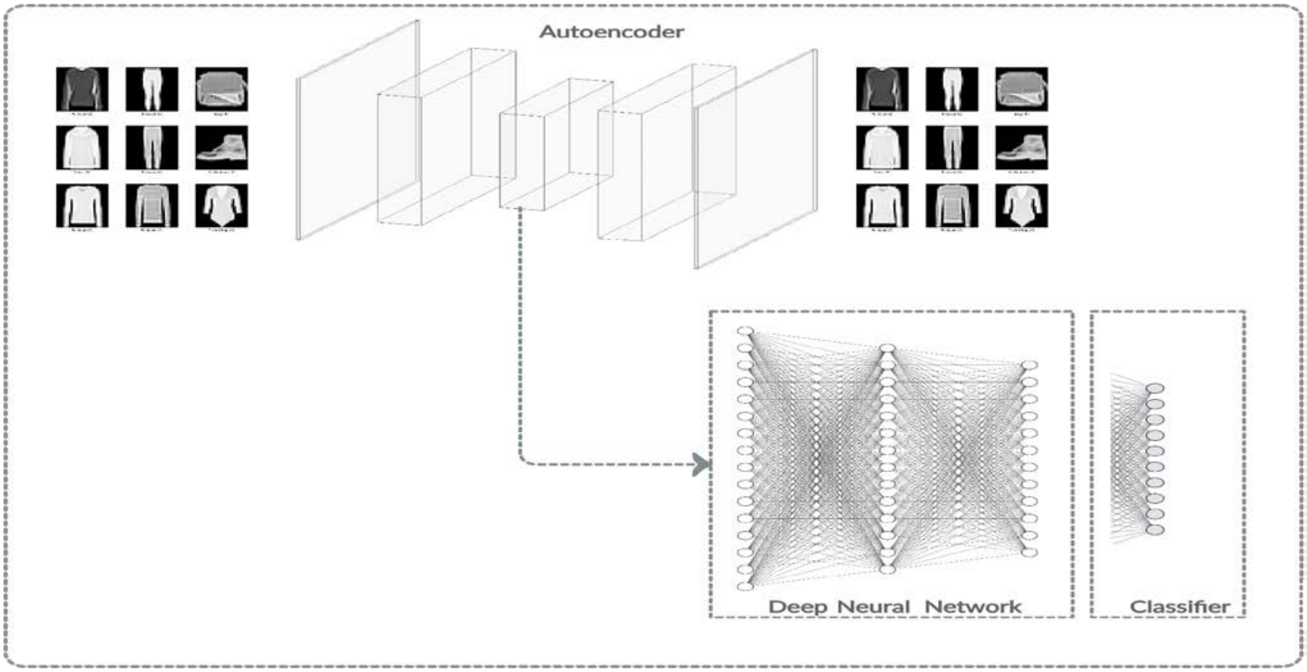


Fig. 1. The Overview Architecture of the Proposed Deep Autoencoder and Deep Neural Networks Framework.

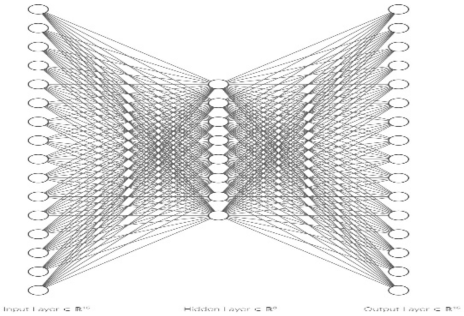


Fig. 2. Typical Autoencoder.

The proposed deep autoencoder consists of decoder and encoder as shown in Table I. The decoder is designed based on ConvNet2D layers followed by BatchNormalization layer and MaxPooling layers, whereas the decoder is based on ConvNet2D followed by UpSampling2D.

B. Deep Neural Networks

Neural networks are a supervised learning that consists of multiple layers to discover underlying feature between different variables as shown in Table III. In this study, deep neural networks take the output of the hidden layer of the spatial autoencoder 7 * 7 layer as an input which was flatten. The proposed deep neural networks consist of eight layers as shown in equation 4.

$$h_n = g\,(W_1 x_1 + b_1) \qquad (4)$$

Furthermore, rectified linear unit activation function (ReLU) is applied to allow our framework to easily get sparse representations [24]. ReLU prunes the negative value to zero, and keeps the positive number x to the same value as shown in equation 5 and Fig. 3.

$$(x) = max(x, 0) \qquad (5)$$

In addition, dropout layer is regularization technique that used to prevent/reduce overfitting issue especially when using huge networks [25, 26]. In our framework, dropout layer is introduced in both autoencoder and deep neural network to avoid the overfitting problem.

C. Classifier

The last layer of this proposed framework is the classifier layer with 10 neurons to classify the output into one of the fashion classes. The softmax activation layer is used to output the probability distribution for categorical classification and defined as follows [27]:

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{k} e^{x_i}} \qquad (x)$$

Where, x_i is the dimension of the input vector, and it will return a value between zero and one for each class and 1 for all classes.

TABLE I. DEEP AUTOENCODER ARCHITECTURE

Specific Parameter	Value/Type
Networks	Decoder and Encoder
Input shape	(28,28,1)
Epochs	15
Batch-size	32
Optimizer	Adam
Loss Function	mean_squared_error
Activation Function	Rectifier Linear Unit (ReLU) And Sigmoid
Pooling	MaxPooling and UpSampling

TABLE II. CLASS NAMES AND EXAMPLE IMAGES IN FASHION-MNIST DATASET

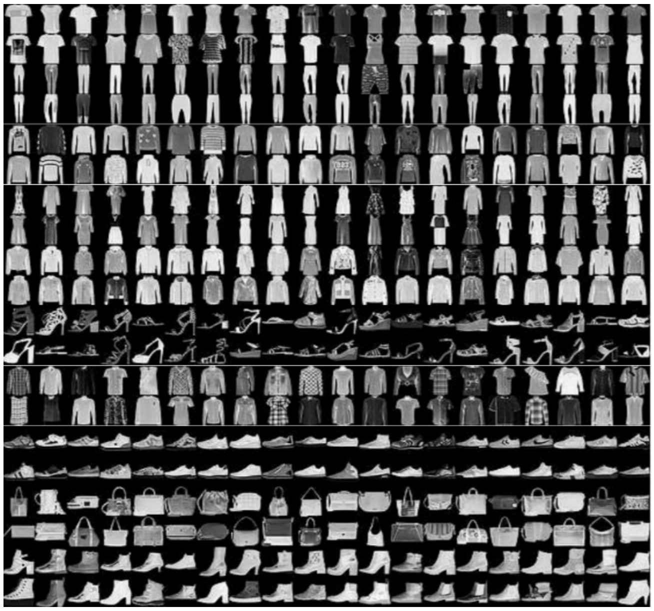
Label	Description	Example
0	T-shirt/top	
1	Trouser	
2	Pullover	
3	Dress	
4	Coat	
5	Sandal	
6	Shirt	
7	Sneaker	
8	Bag	
9	Ankle boot	

TABLE III. DEEP NUERAL NETWORK ARCHITECTURE

Specific Parameter	Value/Type
Layers	7
Input shape	(14, 14, 128)
Epochs	120
Batch-size	64
Optimizer	Adam
Learning Rate	0.001
Loss Function	categorical_crossentropy
Activation Function	ReLU
Dropout	0.5
Last Layer	Softmax

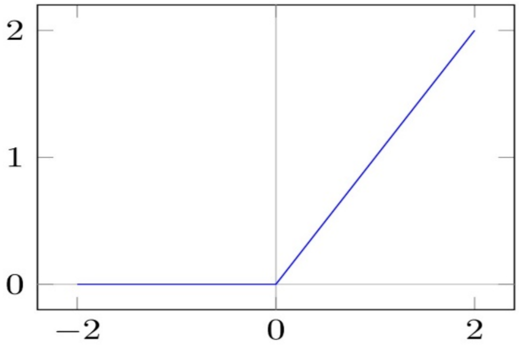


Fig. 3. Rectifier Linear Units (ReLU).

IV. DISCUSSION AND EXPERIMENTAL RESULTS

The goal of this section is to demonstrate the performance of the proposed framework based on deep autoencoder and deep neural networks. The proposed framework is tested on Fashion MNIST dataset. This section is divided into three subsections: (1) Fashion MNIST Dataset, (2) Performance evaluation, and (3) Discussion.

A. Fashion-MNIST Dataset

Fashion-MNIST [28] was introduced in 2017 and considered as replacement of MNIST dataset. This dataset composes of ten categories: t-shirt, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag and ankle boot. Fashion-MNIST consists of 70000 images which divided as follows: 60000 and 10000 for training and test respectively. each category has 6000 and 1000 images for training and test respectively. Images are grayscale with size of 28 by 28 as shown in Table II.

B. Performance Evalution

In this subsection, the classification performance of the proposed Autoencoder framework is evaluated using the Fashion MNIST dataset. We compared the performance of our proposed DeepAutoNN framework with some of the previous proposed approaches in term of overall accuracy. The compared approaches were: Hierarchical Convolutional Neural based on VGG16 and VGG19 [1], VGG11 [12], Convolutional Neural Network with Support vector Machine [14], Convolutional Neural Network followed by Batch Normalization and Skip Connection [7], HitNet [17], Histogram of Oriented Gradient with Multiclass Support Vector Machine [18], Long-Short Term Memory [19], Unconditional and conditional Collaborative Neural Networks [20], Histogram of Oriented Gradient with Local Binary Pattern [21].

As shown in Table IV, the proposed framework achieves the best result with accuracy of 93.4% compared to other proposed approaches.

TABLE IV. PERFORMANCE COMPARISON WITH THE PROPOSED SYSTEM

Model	**Accuracy**
H-CNN using VGG19 [1]	93.33%
VGG-11 [12]	91.5%
CNN-SVM [14]	90.72%
CNN2 + BatchNorm + Skip [7]	92.54%
HitNet [17]	92.30%
HOG + SVM [18]	86.53%
LSTM [19]	88.26%
CNN + U-CoNN [20]	90,61%
CNN + C-CoNN [20]	90,84%
ConvNet [13]	91.00%
HOG + LBP [21]	86.20%
DeepAutoNN-ours	93.36%

Confusion metrics are computed to estimate the classifier performance, and they are calculated as follows.

$$Accuracy\ (Acc) = (TP + TN)/(TP + TN + FP + FN)$$

$$Sensitivity\ (\mathrm{Sen}) = \mathrm{TP} / (\mathrm{TP} + \mathrm{FN})$$

$$Precision\ (\mathrm{Pre}) = TP / (\mathrm{TP} + \mathrm{FP})$$

$$\mathrm{Specificity\ (Spe)} = \mathrm{TN} / (\mathrm{TN} + \mathrm{FP})$$

$$F - Score\ (F1) = (2 \times (Sen \times Pre) / (Sen + Per))$$

Where TP and TN represents the numbers of true-positive results, and true-negative and, respectively, and FN and FP denote the numbers of false-negative and false-positive results, respectively. The performance evaluation of the proposed framework using the evaluation metrics is shown in Table V. Furthermore, confusion matrices are used to further evaluate the multi-classification performance as shown in Fig. 4.

DATASET

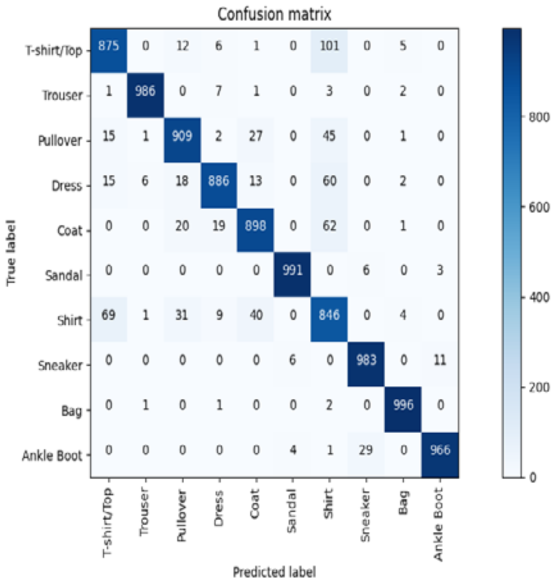


Fig. 4. Confusion Matrix of Classification result with Fashion MNIST.

C. Discussion

In this subsection, the efficiency of the proposed AutoDeep classification framework utilizing fashion images is discussed and analysis. The proposed deep autoencoder is used to extract the hidden and useful features to enhance the classification performance. We found out that using large number of feature maps assess the deep neural networks to converge easily, also, the softmax classifier performs well on the Fashion MNIST. Moreover, increasing the number of hidden layers of neural networks leads to better performance. The proposed framework converges fast as shown in Fig. 5 and Fig. 6 due to the reduce dimensionality using the autoencoder.

The computational time required by DeepAutoNN framework is further analysed which can be divided into three stages: deep autoencoder, deep neural networks and multi-classifier layer. The total time required for the DeepAutoNN framework to process and classify a single image is approximately 0.001 second/image. The proposed DeepAutoNN framework for fashion classification is implemented using Windows 10 with an Intel i7 CPU and 32 gb RAM. TensorFlow, Numpy, sklearn, matplotlib and the Keras library are utilized as tools to implement the framework.

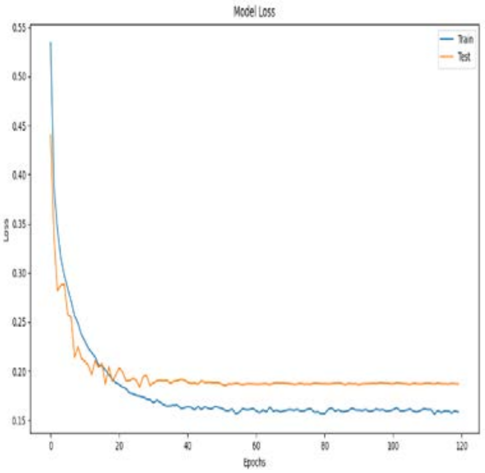


Fig. 5. Training Loss.

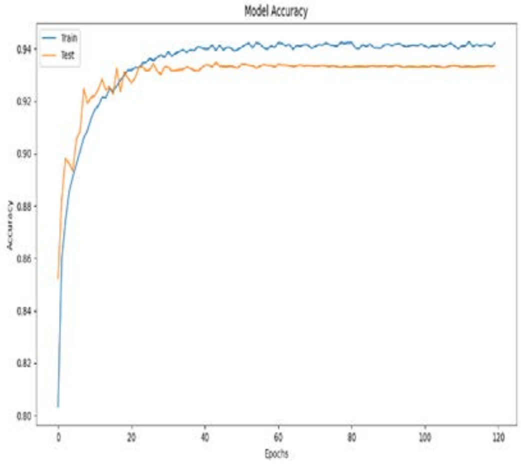


Fig. 6. Training Accuracy.

TABLE V. THE PERFORMANCE EVALUATION USING EVALUATION METRICS

Label	**precision**	**Recall**	**F1-score**
T-shirt/top	90	88	89
Trouser	99	99	99
Pullover	92	91	91
Dress	95	89	92
Coat	92	90	91
Sandal	99	99	99
Shirt	76	85	80
Sneaker	97	98	97
Bag	99	100	99
Ankle boot	99	97	98

V. CONCLUSION

With rapid advancement in deep learning and computer vision techniques, many studies have been conducted on complex vision tasks due to the powerful representation-learning algorithms utilizing advance deep learning techniques. Image detection and classification have been utilized in e-commerce industries such as apparel applications. This paper has proposed a novel framework that can detect and classify fashion images utilizing deep learning techniques. This framework combines a deep autoencoder with deep neural networks to extract the complex patterns and high-level features of fashion images. The framework AutoDeepDNN, demonstrates significant improvements over existing approaches on Fashion MNIST dataset with accuracy of 93.4% %. The future research plan is to explore and investigate generative adversarial networks and its variant to improve the classification results using more fashion datasets.

REFERENCES

[1] Seo, Y. and K.-s. Shin, Hierarchical convolutional neural networks for fashion image classification. Expert Systems with Applications, 2019. 116: p. 328-339.

[2] Pouyanfar, S., et al., A survey on deep learning: Algorithms, techniques, and applications. ACM Computing Surveys (CSUR), 2018. 51(5): p. 1-36.

[3] Dong, Q., S. Gong, and X. Zhu. Multi-task curriculum transfer deep learning of clothing attributes. in 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). 2017. IEEE.

[4] Liu, Z., et al. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[5] Liang, X., et al. Human parsing with contextualized convolutional neural network. in Proceedings of the IEEE international conference on computer vision. 2015.

[6] Hadi Kiapour, M., et al. Where to buy it: Matching street clothing photos in online shops. in Proceedings of the IEEE international conference on computer vision. 2015.

[7] Bhatnagar, S., D. Ghosal, and M.H. Kolekar. Classification of fashion article images using convolutional neural networks. in 2017 Fourth International Conference on Image Information Processing (ICIIP). 2017. IEEE.

[8] Bossard, L., et al. Apparel classification with style. in Asian conference on computer vision. 2012. Springer.

[9] Yamaguchi, K., T.L. Berg, and L.E. Ortiz. Chic or social: Visual popularity analysis in online fashion networks. in Proceedings of the 22nd ACM international conference on Multimedia. 2014.

[10] Han, X., et al. Learning fashion compatibility with bidirectional lstms. in Proceedings of the 25th ACM international conference on Multimedia. 2017.

[11] Nair, P., R. Doshi, and S. Keselj, Pushing the limits of capsule networks. Technical note, 2018.

[12] Duan, C., et al. Image classification of fashion-MNIST data set based on VGG network. in Proceedings of 2019 2nd International Conference on Information Science and Electronic Technology (ISET 2019). International Informatization and Engineering Associations: Computer Science and Electronic Technology International Society. 2019.

[13] Elsaadouny, M., J. Barowski, and I. Rolfes. ConvNet Transfer Learning for GPR Images Classification. in 2020 German Microwave Conference (GeMiC). 2020. IEEE.

[14] Agarap, A.F., An architecture combining convolutional neural network (CNN) and support vector machine (SVM) for image classification. arXiv preprint arXiv:1712.03541, 2017.

[15] Greeshma, K. and K. Sreekumar, Hyperparameter Optimization and Regularization on Fashion-MNIST Classification. International Journal of Recent Technology and Engineering, 2019.

[16] Rajasegaran, J., et al. Deepcaps: Going deeper with capsule networks. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[17] Deliege, A., A. Cioppa, and M. Van Droogenbroeck, Hitnet: a neural network with capsules embedded in a hit-or-miss layer, extended with hybrid data augmentation and ghost capsules. arXiv preprint arXiv:1806.06519, 2018.

[18] Greeshma, K. and K. Sreekumar, Fashion-MNIST classification based on HOG feature descriptor using SVM. International Journal of Innovative Technology and Exploring Engineering, 2019. 8: p. 960-962.

[19] Shen, S., Image Classification of Fashion-MNIST Dataset Using Long Short-Term Memory Networks.

[20] Li, N., Y. Sheng, and H. Ni. CoNN: Collaborative Neural Network for Personalized Representation Learning with Application to Scalable Task Classification. in 2019 International Conference on Computer, Information and Telecommunication Systems (CITS). 2019. IEEE.

[21] Xiao, Z., et al., A Multi-Class Objects Detection Coprocessor With Dual Feature Space and Weighted Softmax. IEEE Transactions on Circuits and Systems II: Express Briefs, 2020. 67(9): p. 1629-1633.

[22] Liu, G., H. Bao, and B. Han, A stacked autoencoder-based deep neural network for achieving gearbox fault diagnosis. Mathematical Problems in Engineering, 2018. 2018.

[23] Wang, W., et al. Generalized autoencoder: A neural network framework for dimensionality reduction. in Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2014.

[24] Glorot, X., A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. in Proceedings of the fourteenth international conference on artificial intelligence and statistics. 2011.

[25] Hinton, G.E., et al., Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580, 2012.

[26] Wu, H. and X. Gu. Max-pooling dropout for regularization of convolutional neural networks. in International Conference on Neural Information Processing. 2015. Springer.

[27] Shen, W. and R. Liu, Tackling Early Sparse Gradients in Softmax Activation Using Leaky Squared Euclidean Distance. arXiv preprint arXiv:1811.10779, 2018.

[28] Xiao, H., K. Rasul, and R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747, 2017.

Adaptive e-Learning AI-Powered Chatbot based on Multimedia Indexing

Salma El Janati[1], Abdelilah Maach[2], Driss El Ghanami[3]
LRIE Laboratory, Mohammadia School of Engineers (EMI)
Mohammed V University, Rabat
Morocco

***Abstract*—With the rapid evolution of e-learning technology, the multiple sources of information become more and more accessible. However, the availability of a wide range of e-learning offers makes it difficult for learners to find the right content for their training needs. In this context, our paper aims to design an e-learning AI-powered Chatbot allowing interaction with learners and suggesting the e-learning content adapted to their needs. In order to achieve these objectives, we first analysed the e-learning multimedia content to extract the maximum amount of information. Then, using Natural Language Processing (NLP) techniques, we introduced a new approach to extract keywords. After that, we suggest a new approach for multimedia indexing based on extracted keywords. Finally, the Chatbot architecture is realized based on the multimedia indexing and deployed on online messaging platforms. The suggested approach aims to have an efficient way to represent the multimedia content based on keywords. We compare our approach with approaches in literature and we deduce that the use of keywords on our approach result on a better representation and reduce time to construct multimedia indexing. The core of our Chatbot is based on this indexed multimedia content which enables it to look for the information quickly. Then our designed Chatbot reduce response time and meet the learner's need.**

Keywords—e-Learning; Chatbot; Speech-To-Text; NLP; Keywords Extraction; Text Clustering; Multimedia Indexing

I. INTRODUCTION

With the rapid evolution of the IT field and the continuous updating of available tools. Learners must undergo continuous trainings in order to improve their skills and ensure technological monitoring [1]. Indeed, training has been strongly affected by the digital transformation. e-Learning (online training) is a perfect example with the digitization of learning. This distance learning technique eliminates the physical presence of a trainer [2]. It has many advantages and is become a part of the learner journey. However, the quantity of multimedia contents offered for the learner increases exponentially, which makes the autonomous learning a complicated task because it is necessary to choose the content adapted to their context to ensure an effective learning [3].

In this context, our study aims to create an indexed database of e-learning multimedia content in order to set up a Chatbot system who offer the appropriate content to each learner according to his or her e-learning needs. The advancements in artificial intelligence and NLP, allowing bots to converse more and more, like real people who share e-learning presentation contents. Predominant Chatbots do not depend exclusively on content, and will frequently appear valuable cards, pictures, joins, and shapes, giving an app-like encounter.

One of the most difficult areas of research is the development of effective Chatbots that emulate human dialogue. It is a difficult task that involves problems related to the NLP (Natural Language Processing) research field [4]. Thanks to NLP techniques and algorithms, it possible to understand the learners' requests based on what the learner is writing. Usually, this task is the core of the Chatbot but there are some limitations as it is not possible to map all learner requests, and current Chatbots do not show remarkable performance due to the unpredictability of the thinking of the learner during a conversation [5]. One of the most important task in setting up Chatbot is the design of the conversational flow. In fact, we suggest a new approach for a successful conversation based on keywords extractions, which it is important to handle with all learners requests and provide the adequate content.

This paper is structured as follows: the second section is a related work. In the third section, we will present the suggested approach. Then, the fourth part will be devoted to our result and our simulation. Finally, we will conclude with a recommendation approach to improve the suggested approach.

II. RELATED WORK

The use of the advanced technology in data science enables to improve the quality of e-learning content. In this paper, we suggest a framework that uses Natural Language Processing (NLP) and Keywords Extraction as a chatbot engine for e-learning. Thus, in this section we will review the literature related to chatbot and Automatic Keywords Extraction (AKE).

A. Chatbot

Chatbots are a virtual assistant capable of chatting with users and responding to their requests. They are increasingly using speech synthesis techniques to produce their messages as the user types their interventions into a text field on the web page.

The design of a Chatbot to meet the needs of users was always a concern in the field of information retrieval [4]. Depending on how Chatbots are programmed, we can divide them into two large groups: those that are programmed according to predefined commands (rule-based Chatbot) [5] and those based on artificial intelligence (AI) [6].

AI Chatbots using machine learning are designed to understand the context and intent of a question before formulating an answer. There are two types of AI Chatbots: Generative Chatbots [7] and Information Retrieval Chatbots [4]. Fig. 1 presents Chabot's types.

The generative Chatbot is built in order to be able to act to any context or interlocutor and also in nonprogrammed situations. Such a conversational agent relies on the new artificial intelligence techniques such as deep learning and neural network to generate his responses word by word [8]. Thus, these bots can construct answers to users questions themselves. The problem with this approach is that it is too general and Chatbots struggle to have a coherent conversation or even to produce syntactically valid sentences with current models.

The Information retrieval Chatbots are the chatbot adapted to a given context, which builds its responses using a set of sentences that have been given to it in advance. These chatbots are based on retrieving information from the user's question and seek the most suitable answer using NLP [9]. This type of Chatbots is best suited for closed domain systems. This approach guarantees grammatically correct answers and simplifies the learning task for the algorithm, because it allows constructing the model on a training data set smaller than the big amount of data required in the case of a generative Chatbot [4].

Thus, in our case study we chose to design a Chatbot based on retrieval information since it allows to exploit the results obtained from videos indexing and also have some control over the responses generated by the Chatbot.

Our contribution consists of using a keyword extraction technique adapted to the multimedia e-learning content we have. Thus, the Chatbot will be based on the keywords instead of all the text in order to find the adequate answer to the learner's need in a fast and efficient way. The next section will be dedicated to keyword extraction techniques found in literature.

B. *Automatic Keywords Extraction (AKE)*

Keyword extraction involves identifying the words and phrases representing the main subjects of a document. High quality keywords can make it easier to understand, organize and access the content of the document. AKE from a document has been used in many applications, such as information retrieval [10], text synthesis [11], text categorization [12], and opinions mining [13].

Most existing keyword extraction algorithms address this problem in three steps (Fig. 2): First, the candidate keywords (i.e. words and phrases that can be used as keywords) are selected in the content of the document. Second, candidates are either classified using a candidate weighting function (unsupervised approaches) or classified into two classes (keywords / no keywords) using a set of extracted characteristics (supervised approaches). Third, the most weighted first N candidates with the highest confidence scores are selected as keywords [14].

In this section, we provide an overview of keyword extraction methods. Our goal is not to detail the functioning of these methods, but rather to have an overview of their basic principle and their classification in order to identify the suitable methods for our case study.

1) Statistic approach: Statistical Approaches is considered one of the simplest techniques used to identify keywords within a text. These approaches do not require training data in order to extract the most important keywords in a text. it seeks to define what a keyword, based on certain statistical features and study their relation with the notion of importance of a candidate term. the more the term candidate is considered important in analysed document, and the more it will be relevant as a keyword.

TF-IDF [15] and Likey [16] are two methods, which compare the behavior of a candidate term in the analysed document with its behavior in a collection of documents (reference corpus). The objective is to find candidate terms whose behavior in the document varies positively compared to their overall behavior in the collection. In both methods, this is expressed by the fact that a term is important in the document analysed if it is largely present, when it is not in the rest of the collection.

Yake! Approach [17] focuses on statistical features that do not require external dictionaries and look on characteristics, which can be calculated using only current document. These approaches are based on characteristics such as the position of the first occurrence of a candidate, the word frequency, the case, and the frequency with which a word appears in different sentences.

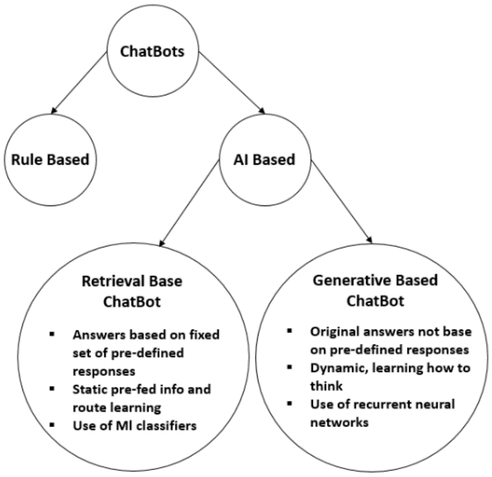

ChatBots
Rule Based
AI Based
Retrieval Base ChatBot
Answers based on fixed set of pre-defined responses
Static pre-fed info and route learning
Use of Ml classifiers
Generative Based ChatBot
Original answers not base on pre-defined responses
Dynamic, learning how to think
Use of recurrent neural networks

Fig. 1. Chatbots Types.

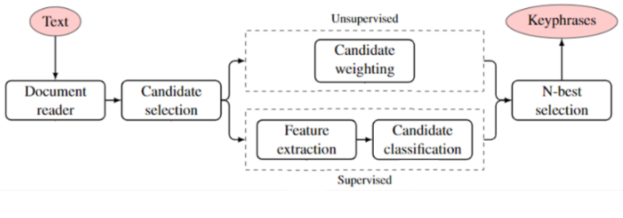


Fig. 2. General approach for Automatic Keywords Extraction.

2) Graph based approach: Graph-based approaches consist of representing the content of a document as a graph. The methodology applied comes from PageRank [18], an algorithm for ranking web pages (graph nodes) based on the recommendation links that exist between them (graph edges).

TextRank [19] and SingleRank [20] are the two basic adaptations of PageRank for AKE. In these, web pages are replaced by text units whose granularity is the word and an edge is created between two nodes if the words they represent co-occur in a given word window. When running the algorithm, a score is associated with each candidate keyword, which represents the importance of the node in the graph.

To improve TextRank / SingleRank, Liu et al. suggest a method, which aims to increase the coverage of all the key terms extracted in the analysed document (TopicalPageRank) [21]. To do this, they try to refine the importance of words in the document by taking into account their rank in each topic. The rank of a word for a topic is obtained by integrating into its PageRank score the probability that it belongs to the topic. The overall rank of a candidate term is then obtained by merging its ranks for each topic.

3) Word embedding approach: Word embedding [22] is a new way of representing words as vectors, typically in a space of a few hundred dimensions. A word is transformed into numbers. The word representation vector is learned based on an iterative algorithm from a large amount of text. The algorithm tries to put the vectors in space in order to bring together the semantically close words, and to move away the semantically distant words. By finding the closest words in the embedding space of a given word as input, the model identifies synonyms or intruders in a list of words. Once a model is obtained, several standard tasks become possible.

Different methods based on the word embedding are suggested for representing entire documents or sentences [23]. Skip-thought [24] provides sentence embedding trained to predict neighbouring sentences. Sent2Vec [25] generate sentence embedding using word n-gram representation.

With words embedding the keyword can be extracted using the cosine similarity in the word space representation. EmbedRank [26] is a word embedding based approach to automatically extract key phrases. In this method, documents or word sequences of arbitrary length are embedded into the same feature space. This enables computing semantic relatedness between document and candidate keyword by using the cosines similarity measures.

4) Supervised approach: Supervised approaches are methods who able to learn to perform a particular task, in this case the extraction of keywords. Learning is done through a corpus whose documents are annotated in keywords. The annotation allows to extract the examples and counter-examples whose statistical and / or linguistic features are used to teach a binary classification [27]. These classifications consist of indicating if a candidate term is a keyword or not. Many supervised algorithms are used in various fields. They can adapt to any task, including AKE task. Algorithms used for this construct probabilistic models, decision trees, Support Vector Machine (SVM) or even neural networks.

KEA is a method which uses a naive Bayesian classification to assign a likelihood score to each candidate term, the aim being to indicate whether they are keywords or not [28]. These approaches use three conditional distributions learned from the learning corpus. The first is the probability that each candidate term is labelled yes (keyword) or no (no keyword). The other two stand for two different statistic features which are the TF-IDF weight of the term candidate and its first position in the document.

Nguyen et al. propose an improvement (WINGNUS) [29] by adding a set of features such as: First and last occurrences (word off-set), Length of phrases in words, whether a phrase is also part of the document title, number of times a phrase appears in the title of other document. Adding these features improves the performance of the original version of KEA, but only when the amount of data is large enough. Suggested Architecture.

III. Suggested Approach

In this section, we suggest an approach of an adaptive Chatbot based on Retrieval information. This approach consists of indexing e-learning multimedia using keyword extraction. These indexed contents will be integrated in the Chatbot engine in order to offer the adequate content to the learner's needs.

The suggested approach is based on four steps, which will be detailed:

1) Extracting metadata from the e-learning database.
2) Speech to text Processing.
3) Automatic Keywords Extraction.
4) Suggested Chatbot design.

The first step of our approach consists of analyzing e-learning multimedia content and extracting as much information as possible from e-learning content in order to build our database. The second step will be devoted to standardize the e-learning content by transforming all content to text, so we use the Speech To Text techniques in order to extract text from all multimedia content. In the third step, we will suggest methods for extracting keywords from the extracted text. We will test several approaches and suggest a new approach adapted to our problematic.The last step describe chatbot design based on keywords resulting from the previous steps. Fig. 3 summarizes the methodology of our approach.

A. Extracting Metadata from e-Learning Database

In our study, we use a various source of e-learning content in order to construct database, which will be the base of our Chatbot recommendation. e-Learning sources provide content in different types of information like video, speech and, text. The first step of our approach consists of extracting as much information as possible from e-learning content in order to build our database.

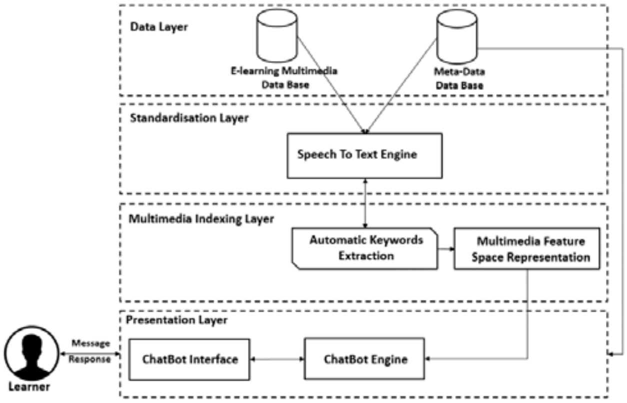


Fig. 3. Methodology Flow Chart.

In order to extract metadata from e-learning multimedia, we design a Python script that automatically retrieves metadata. This data frame is rich in information, it mainly contains: ID, the title, the category, the description, the subtitle if it is +available, the date of the online publication, and the author. In addition, the metadata contains the audio of multimedia if it is a video content. This metadata information is used in the Chabot engine in order to organize multimedia e-learning content and to give an easy content access. Fig. 4 shows the metadata extraction script.

The second step of our approach is to standardize the e-learning content by transforming all types into text (video to text and speech to text). The next sub-section is dedicated to describe the speech to text approach used in our case study.

B. Speech to Text Processing

Automatic Speech Recognition (ASR) is the process by which speech is transcribed into text. This technology has many useful applications ranging from hands-free car interfaces to home automation. Although speech recognition is an easy task for humans, it has always been difficult for machines. Since 2012, the use of Deep Neural Networks (DNN) has considerably improved the accuracy of speech recognition [30].

Deep Learning-based ASR systems today have achieved even better results than humans in languages like English [Baidu's Deep speech 2] [31]. These advances have been propelled by the use of large amounts of data (up to tens of thousands of hours of transcribed speech) and by an enormous parallel computing power controlled by GPUs.

The general approach of ASR systems based on Deep Learning and which is currently used by almost all systems offered by large groups such as google, IBM ... This approach follows the architecture presented in Fig. 5.

This approach generally starts from converting audio into a feature matrix to feed it into the neural network. This is done by creating spectrograms from audio waveform. The spectrogram input can be considered as a vector at each timestamp. A 1D convolutional layer extricates highlights out of each of these vectors to provide a grouping of highlight vectors for the LSTM layer to handle. The (Bi) LSTM output layer is passed to a Fully Connected layer for each time step which gives a probability distribution of the character at that time step using softmax activation.

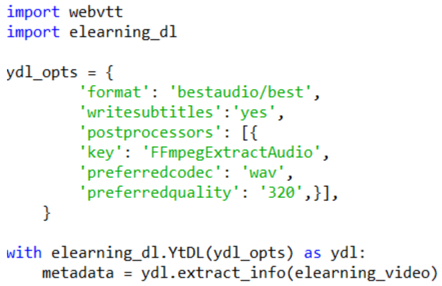
The problem of automatically recognizing speech with the assistance of a computer is a difficult task, due to the complexity of the human dialect. To solve this problem, several challenges must be addressed: Microphone poor quality, background noise, speaker variability and so on. All these possibilities must be included in the training step for Deep Network to process them. Thus, to create a voice-recognition system that achieves the performance of Siri, Google Now or Alexa, it is mandatory to have a large amount of data for the training step.

Given the difficulty of acquiring a massive database to train our own ASR system, we chose in our study to evaluate the APIs that are available on the market. Thus, we develop a Python Script that implement different Speech Recognition API's using 'Speech Recognition' library.

In order to compare different methods proposed by API's for automatic transcription, we need to evaluate the performance of each model. The key metric for transcription is accuracy. How closely the words within the created transcript match the talked words within the unique sound.

To calculate the accuracy of the automatic transcription, we will use the metric Word Error Rate (WER). The WER is a very simple and widely use measure for transcription accuracy. It is a number, calculated as the number of words needed to be inserted or changed or deleted to convert the transcript hypothesis into the reference transcript, divided by the number of words in the reference transcript. (It's the Levenshtein distance for words, measuring the minimum number of single words edits to correct the transcription.) A perfect match has a WER of zero; larger values indicate lower accuracy and thus more editing.

$$Word\ Error\ Rate = \frac{Insertions + Deletions + Substitutions}{Number\ of\ Words\ in\ Reference\ Transcript}$$



```
import webvtt
import elearning_dl

ydl_opts = {
        'format': 'bestaudio/best',
        'writesubtitles':'yes',
        'postprocessors': [{
        'key': 'FFmpegExtractAudio',
        'preferredcodec': 'wav',
        'preferredquality': '320',}],
    }

with elearning_dl.YtDL(ydl_opts) as ydl:
    metadata = ydl.extract_info(elearning_video)
```

Fig. 4. Metadata Extraction Script.

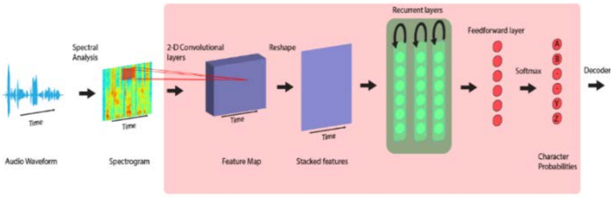


Fig. 5. Global Architecture of ASRs based on Deep Learning.

Fig. 6 and Table I show WER evaluation obtained from the APIs transcription compared to ground truth.

The results obtained show that the transcription of YouTube is the most efficient, followed by the transcription of Houndify. This is because the YouTube transcription is generally based on the publisher's recommendation, but this transcription is not always available. Houndify API specializes in detecting lyrics in music and that can explain the good performance of this API, since our e-learning data source mainly contains a speech with a music background. We can also notice that IBM and Google services have a similar performance since they are based on similar resources for learning phase. Finally, the Wit transcription is the least efficient and this is generally due to resources for the learning phase which are quite limited.

Based on results obtained is this step, we decided to make the approach shown on Fig. 7, in order to normalize multimedia e-learning types and extract text from video and audio e-learning content.

Architecture shown on Fig. 7 contains three different processing; each one is adapted for one type of e-learning multimedia:

- Video Processing: Extract plain text from original video if it is provided by the author, otherwise apply Speech to Text API in order to extract text from video speech.
- Speech processing: Apply Speech Recognition in order to extract text from e-learning speech.
- Text processing: Extract plain text from e-learning text.

This architecture enables to have a normalized Database, which contains plain text for every e-learning multimedia sources. The next step is to extract from the plain text the important words (keywords) that will represent each multimedia content.

C. Automatic Keywords Extraction

In this subsection, we will present a keyword extraction evaluation based on the techniques described in our literature review.

1) Keyword extraction approach evaluation: In order to evaluate the performance of the different algorithms, a first approach consists of applying a manual evaluation based on human judgement to decide whether the keywords are representatives of the content of a document or not [17]. Nevertheless, manual evaluation of Automatic Keywords Extraction is difficult and time consuming.

Researchers have therefore developed some automatic evaluation systems based on partial correspondence. Automatic key phrase extraction methods have generally been evaluated based on the number of N first candidates which correspond correctly to the reference keywords. This number is then used to calculate the accuracy, recall and F-score for a set of keywords.

In the same perspective, we will compare keywords obtained by each approach with the set of Tags provided by the author of some multimedia content and which we consider to be the ground truth. So, we will use the F1 score which is based on precision and recall. Precision is defined as the number of correctly predicted keywords out of the number of all predicted keywords, and recall is defined as the number of correctly predicted keywords out of the total number of keywords in the ground truth set. Note that, to determine the correspondence between two key words, we use Porter Stemmer for the preprocessing in order to consider keywords which have the same root word.

Then, we calculated F1 for the first 20 keywords on our set of e-learning multimedia. Fig. 8 represents box plot graphs for evaluated algorithms as well as Table II presents the statistical measures (Min, Max and the mean).

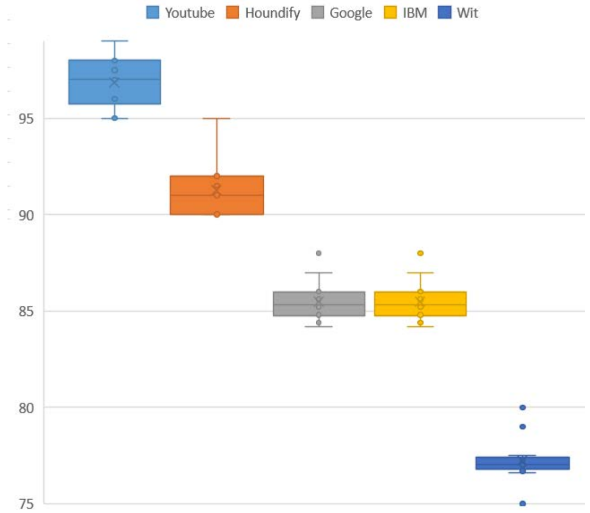


Fig. 6. WER Index of Transcription obtained from the ASR APIs.

TABLE I. WER INDEX OF TRANSCRIPTION OBTAINED FROM ASR APIS

	YouTube	Houndify	Google	IBM	Wit
WER	97,04	91,12	86,83	86,13	75,23

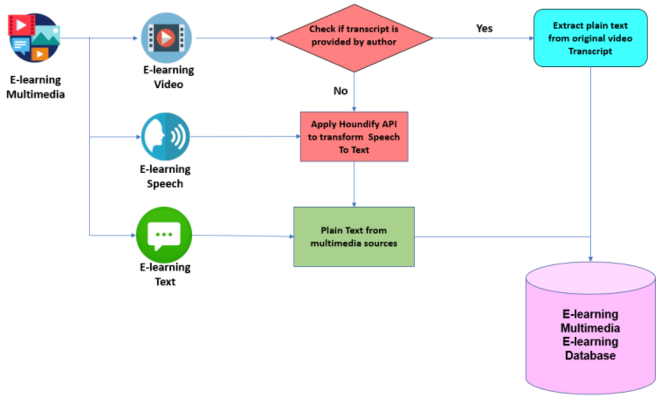

E-learning Multimedia
E-learning Video
E-learning Speech
E-learning Text
Check if transcript is provided by author
Yes
No
Extract plain text from original video Transcript
Apply Houndify API to transform Speech To Text
Plain Text from multimedia sources
E-learning Multimedia E-learning Database

Fig. 7. Architecture of the approach chosen for Multimedia Normalization (Video/Speech to Text Normalization).

Based on Fig. 8, we note that the F1-score of all approaches shows a very large variability. Thus, the relevance of the generated keywords is not stable enough. We can deduce that all approaches evaluated are not adapted to all the content of our e-learning database, and that each approach generates a set of relevant keywords only for a part of our e-learning corpus. In this context, we suggest a new framework based on ensemble methods in order to improve keywords consistency.

2) Suggested framework for automatic keyword extraction: In order to propose a framework adapted to our case study, our main idea consists of suggest an approach allowing to combine the results obtained from all approaches by using a voting system.

Basically, the first step our approach is to apply all the AKE methods in order to obtain the first N keywords with their weights according to each method.

After obtaining the keywords with their weights for each method, the second step of our approach is to normalize the weight of each keyword by dividing its weight for each method by the weight of the keyword with the maximum weight. Thus, the first keyword will have a weight equal to 1.

$$w_{ij} = \frac{w_{ij}}{\max\limits_{i}(w_{ij})} \ \forall\, i \ \in \{keywords\}, \forall\, j \in \{models\}$$

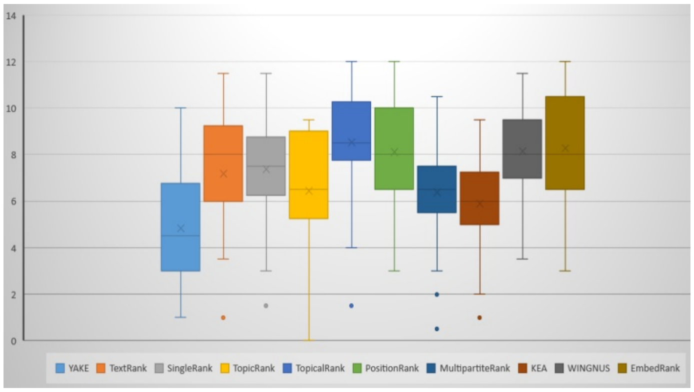


Fig. 8. Box Plot Evaluation Graph for each Keyword Extraction Approach.

TABLE II. WER INDEX OF TRANSCRIPTION OBTAINED FROM ASR APIS

	Min	Average	Max
YAKE	1	4,82	10
TextRank	1	7,18	11,5
SingleRank	1,5	7,38	11,5
TopicRank	0	6,44	9,5
TopicalRank	1,5	8,54	12
PositionRank	3	8,1	12
MltipartitieRank	0,5	6,36	10,5
KEA	1	5,88	9,5
WINGNUS	3,5	8,14	11,5
EmbedRank	3	8,26	12

Then, a global weight for each keyword is calculated by summing the normalized weights. The normalization step avoids favoring one method over another and thus allow to obtain a global weight based on a non-discriminatory vote. Also, to avoid generate similar keywords, the weights of keywords with the same root word are grouped together by considering them as a single candidate keyword.

$$w_i = \sum_{i^{ème}\, keyword\ in\ keywords\, model} w_{ij} \forall\, i \in \{keywords\}$$

Finally, candidate keywords are ranked based on their global weight and the top N are chosen as keywords to represent each e-learning multimedia content.

We notice that the choice of the models to be considered in order to build the global weight of each candidate keyword is important. Thus, to set up an adequate approach to our case study, several configurations are evaluated, namely:

- Voting system based on all methods (Keyword Vote).
- Voting system based on one method per approach (statistic, graph based, word embedding, supervised approach) (Keyword Vote 2).
- Voting system based on best methods (PositionRank, Topical Rank, EmbedRank, Wingnus) (Keyword Vote ++).

Fig. 9 represents a comparison between different configurations of our approach with the previous methods. The results obtained confirm that the proposed approach (Keyword Vote) enable to obtain results with very little variability. Thus, all the keywords generated are relevant for most of the e-learning content. In addition, Fig. 9 shows that (Keyword Vote ++) achieves the best performance. This is due to a good combination of models which stabilizes the generation of relevant keywords.

Based on the results obtained, we can deduce that Keyword Vote ++ makes allow generating a set of relevant and diversified keywords in order to represent each e-learning content. This approach is chosen to extract the keywords from the text of each multimedia content.

3) Suggested chatbot design: To design our chatbot we suggest an architecture composed with two main parts. The first part consists in designing the Chatbot Backend which contains the NLP engine allowing to understand the intention of the user and to propose the most adequate response to the needs of the learner. The second part consists in developing the Chatbot Frontend which constitutes the interface allowing the user to interact with the Chatbot. Fig. 10 shows the overall Chatbot architecture.

a) Chatbot BackEnd: The main role of the Chatbot Backend is to deal with the user's question using NLP Engine and then offer most similar answers to the learner's needs. So, to design the Chatbot Backend we use the approach proposed which is based on extracting keywords using Keyword Vote ++. Fig. 11 shows the backend architecture of the chatbot.

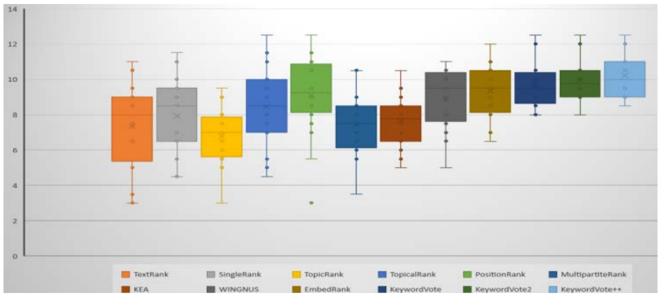


Fig. 9. Box Plot of comparison between different Configurations of our approach with the previous methods of Keyword Extraction.

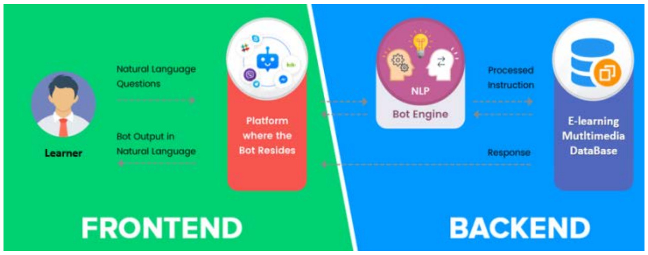

Learner
Natural Language Questions
Bot Output in Natural Language
Platform where the Bot Resides
NLP
Bot Engine
Processed Instruction
Response
E-learning Mutltimedia DataBase
FRONTEND
BACKEND

Fig. 10. Chatbot Design.

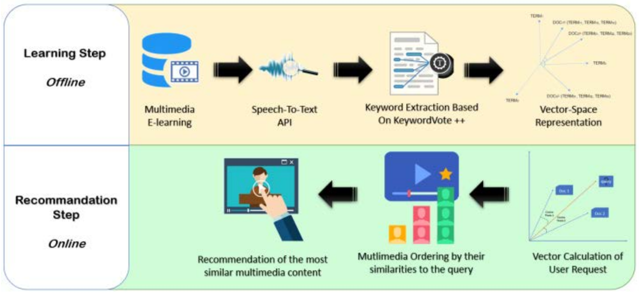

Learning Step
Offline
Multimedia E-learning
Speech-To-Text API
Keyword Extraction Based On KeywordVote ++
Vector-Space Representation
Recommandation Step
Online
Recommendation of the most similar multimedia content
Mutlimedia Ordering by their similarities to the query
Vector Calculation of User Request

Fig. 11. Backend Chatbot Architecture.

The backend process consists of two steps. In the first step, the keywords proposed approach is used to extract the relevant keywords from the e-learning content, then building the representation space of each multimedia e-learning using the terms/documents matrix. This first step is carried out in offline mode since it is not linked to user demand and allows the multimedia indexing. The second step is triggered at each user instruction and consists of calculating the representation of the query in the multimedia indexing space, then recommending to users the most similar content to his request.

b) Chatbot FrontEnd: The Chatbot Frontend consists of the interface design that allows to receive the user instructions and interact with the Backend to display the appropriate response. Thus, to ensure interaction with the user, the Chatbot must have a user-friendly interface. In fact, the Chatbot interface is based on messaging platforms to interact with the user.

In order to develop the Frontend of our Chatbot, we will use the design tools offered by messaging platforms. Our choice was towards the messenger Bot API offered by Slack [32] since on the one hand, it provides the possibility of interacting with a Python script which allows us to set up our approach developed in the Backend, and, on the other hand, Messenger is the most messaging platform used by learners to interact with each other. Thus, implementing our e-learning Chatbot on Slack Messenger has the advantage of offering Chatbot services to learners in an interface that they are used to, which will improve the user experience.

IV. Results and Simulation

A. Results

The previous chapter explained the suggested approach for designing the e-learning chatbot. Indeed, our chatbot is based on a feature space for representing multimedia content built from the most relevant keywords instead of using the overall text which is extracted from multimedia content.

In order to show the advantage of the proposed approach, we apply it on a very large corpus with predefined categories. Indeed, we will compare two approaches:

The approach which is based on the construction of the terms / documents matrix by calculating the weight of the TF-IDF for all words in document [15].

The proposed approach which is based on the Keyword Vote ++ algorithm to extract the keywords and then use the list of keywords to build the similarity matrix between the documents by calculating TF-IDF weight for just the relevant keywords.

Both approaches provide a similarity matrix which will be used to obtain a hierarchical clustering of the multimedia content. In order to validate the clusters obtained from the two approaches, we use the known categories of the corpus as a ground truth. Thus, given the knowledge of class assignments from the ground truth, it is possible to define an intuitive evaluation measure using conditional entropy analysis. Then, we measure two score that aims to identify the Homogeneity and Completeness of each clustering assignment:

- Homogeneity: mean that all members with the same cluster belong to the same class.

- Completeness: all the members of a given class are in the same cluster.

The two concepts are between 0 and 1. Thus, on the basis of its two score another measure called V-measure can be calculated. Indeed, V-measure is the harmonic mean of the two scores.

To complete our evaluation, we use the Rand Index. It measures the similarity between two partitions. Rand's index measures the percentage of correctly classified decisions and is defined as follows:

$$RI = \frac{a+b}{a+b+c+d}$$

Let X: the partition obtained from the clustering algorithm and G: the partition obtained from the ground truth. So:

- $a:$ the number of element pairs that is in the same cluster in X and the same cluster in G.

- $b:$ the number of element pairs that is in a different cluster in X and a different cluster in G.

- c: the number of element pairs that is in the same cluster in X but in different clusters in G.
- d: the number of element pairs that is in a different cluster in X but in the same cluster in G.

Fig. 12 illustrates the results obtained by applying the two approaches on our corpus.

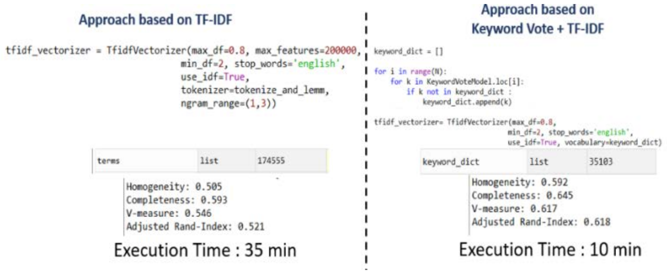
Based on the results obtained, we can confirm that the proposed approach reduces the representation space and thus makes it possible to represent documents with only 35103 terms instead of 174 555 terms (a reduction of the dimension of 80%). It also reduces the time required for the construction of the dendrogram which was reduced at 10 min instead of 35 min (a 70% decrease in execution time). Regarding the validation indices, we notice that there is a remarkable improvement in indices with an 8% gain in performance. This confirms that selecting terms based on keywords helps create create a more homogeneous tree structure for indexing multimedia content and with less execution time.

B. *Simulation*

In order to show the suggested chatbot in action, we propose a simulation based on different multimedia e-learning content which concerns tutorial of different Business Intelligence and Artificial Intelligence Tools. The first step is to apply our suggested Keyword Extraction approach in order to construct an indexed e-learning content. Fig. 13 show Dendrogram obtained by our approach.

The dendrogram provides a hierarchical representation of our database of e-learning contents. Indeed, we notice that multimedia contents are grouped in a way that we can easily distinguish between the different categories contained in our e-learning database. This confirms that the proposed approach is well suited to our case study. The indexed multimedia representation allows us to organize our database in order to facilitate information access and offer content adapted to each user. Indeed, this indexed database will be used into chatbot engine to response to the learner's needs. Our Chatbot design allows to interact with learners in two ways: Quick Reply and Carousel.

1) Quick reply: Quick reply allows to create short instant responses that can be selected by users. Indeed, we use this form to create suggestions of course categories to the user. Fig. 14 shows an example of quick reply.



Approach based on TF-IDF

```
tfidf_vectorizer = TfidfVectorizer(max_df=0.8, max_features=200000,
                                   min_df=2, stop_words='english',
                                   use_idf=True,
                                   tokenizer=tokenize_and_lemm,
                                   ngram_range=(1,3))
```

terms	list	174555

```
Homogeneity: 0.505
Completeness: 0.593
V-measure: 0.546
Adjusted Rand-Index: 0.521
```

Execution Time : 35 min

Approach based on Keyword Vote + TF-IDF

```
keyword_dict = []

for i in range(N):
    for k in KeywordVoteModel.loc[i]:
        if k not in keyword_dict :
            keyword_dict.append(k)

tfidf_vectorizer= TfidfVectorizer(max_df=0.8,
                                  min_df=2, stop_words='english',
                                  use_idf=True, vocabulary=keyword_dict)
```

keyword_dict	list	35103

```
Homogeneity: 0.592
Completeness: 0.645
V-measure: 0.617
Adjusted Rand-Index: 0.618
```

Execution Time : 10 min

Fig. 12. Results comparison of the Proposed Approach with the Classic Approach for Creating a Tree Structure.

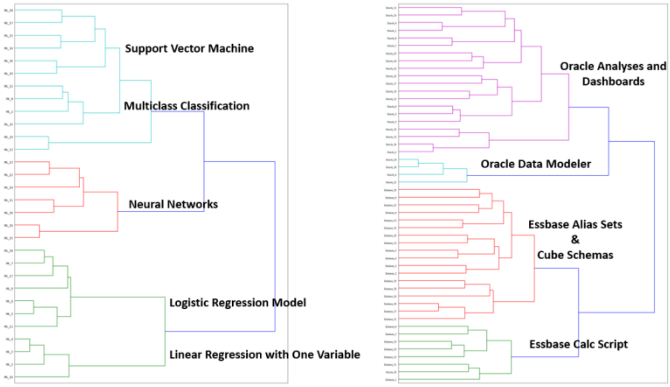
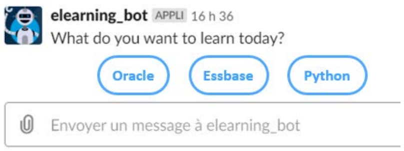


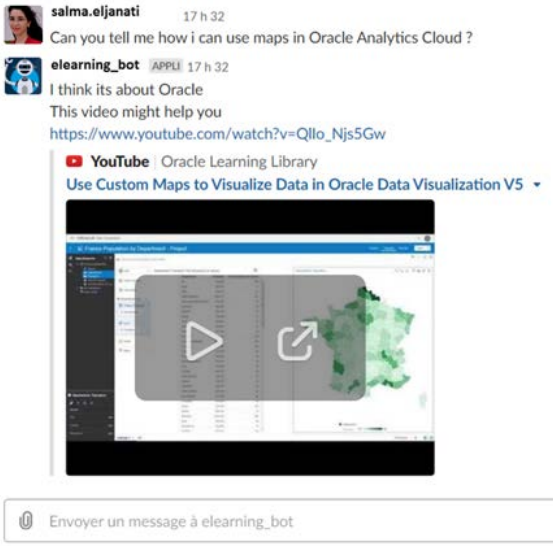
Fig. 13. Hierarchical view resulted from Multimedia Indexing.



elearning_bot APPLI 16 h 36
What do you want to learn today?

Fig. 14. Messenger Bot Quick Reply Example.

2) Carrousel: Another way to display results to the learner is carousel. A carousel is used when a lot of data must be presented to the learner. The buttons that accompany this form can either return a personalized message to the bot as a specialized command to trigger a flow or redirect to a URL. We use this form to recommend multimedia content which fit the user request. Fig. 15 illustrates an example of using carousel in our Chatbot.



salma.eljanati 17 h 32
Can you tell me how i can use maps in Oracle Analytics Cloud ?

elearning_bot APPLI 17 h 32
I think its about Oracle
This video might help you
https://www.youtube.com/watch?v=Qllo_Njs5Gw

YouTube | Oracle Learning Library
Use Custom Maps to Visualize Data in Oracle Data Visualization V5

Fig. 15. Carrousel example.

V. Recommendation

The proposed ChatBot design allows the learner to offer the multimedia content adapted to their needs. However, its functionality is limited and does not allow it to interact effectively in the case of general questions by the learner. One of the ways to improve ChatBot is to integrate a Chit-Chat to simulate human conversation.

Fig. 16 shows the new architecture proposed that will be the aim for our future work.

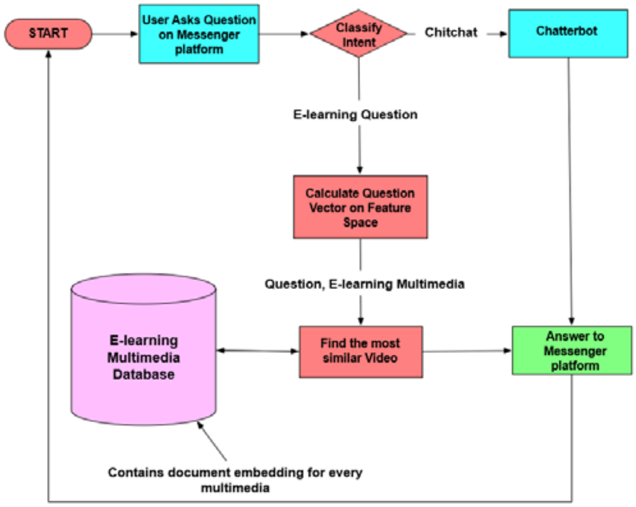

START
User Asks Question on Messenger platform
Classify Intent
Chitchat
Chatterbot
E-learning Question
Calculate Question Vector on Feature Space
Question, E-learning Multimedia
E-learning Multimedia Database
Find the most similar Video
Answer to Messenger platform
Contains document embedding for every multimedia

Fig. 16. Architecture for Integrating Chit-Chat into the Design of ChatBot.

After receiving the question from the user, a classifier will predict the class of the question and determine whether the question is related to the area of e-learning or a general question. Then, depending on the class category, the ChatBot offers an answer according to two scenarios:

- Offer the adapted e-learning multimedia using the approach proposed in this paper.
- Suggest a response from the Chit-Chat database The integration of this module makes the conversation as natural as possible.

VI. Conclusion

The aim of our study is to develop a Chatbot allowing the interaction with learners and suggest the e-learning multimedia content, which fit their learning needs. To achieve this objective, first we set up an analysis of e-learning contents in order to extract the maximum amount of information. Indeed, we propose an approach based on the Speech-To-Text APIs to extract text from different sources of multimedia.

Based on the information obtained from the extracted text, we set up the step of keywords extraction. In this step, we evaluate different algorithms proposed in the literature. Then, we conclude that they are not suitable for all multimedia contained in our e-learning database. Thus, we propose a new approach making it possible to combine the results of different approaches using a voting system. After that, we proceed to indexing of the e-learning content by constructing a tree structure allowing the organization of the information and facilitating the access to the e-learning content.

Finally, we design the ChatBot core which was divided into Backend / Frontend. On the one hand, the Backend design is mainly based on the proposed approach to indexing e-learning multimedia content and which constitutes the engine of the NLP used. On the other hand, the design of the Frontend is based on Slack Messenger platform which offers an interface facilitating interaction with learners.

Our methodology aims to have an efficient way to represent the multimedia content based on keywords. The use of keywords on our approach result on a better representation and reduce time to construct multimedia indexing. The core of our chatbot is based on this indexed multimedia content which enables it to look for the information quickly. Then our designed chatbot reduce response time and meet the learner's need.

The proposed Chatbot design allows the learner to get the multimedia adapted to their needs. However, its functionality is limited and does not allow it to interact effectively in the case of general questions by the learner. Our future work will focus on integrating a Chit-Chat to simulate a human conversation, also we will integrate voice recognition on the chatbot in order to enlarge the scope of the chatbot interactions.

References

[1] El Janati, S., Maach, A., El Ghanami, D., “Learning Analytics Framework for Adaptive E-learning System to Monitor the Learner’s Activities,” International Journal of Advanced Computer Science and Applications, vol. 10, no. 8, 2019.

[2] El Janati, S., Maach, A., El Ghanami, D., “SMART education framework for adaptation content presentation”. Procedia Computer Science, 2018, vol. 127, p. 436-443.

[3] El Janati, S., Maach, A., El Ghanami, D., “Context aware in adaptive ubiquitous e-learning system for adaptation presentation content”. Journal of Theoretical and Applied Information Technology, 2019, 97(16), pp. 4424-4438.

[4] Yan, Z., Duan, N., Bao, J., Chen, P., Zhou, M., Li, Z., & Zhou, J., “Docchat: An information retrieval approach for chatbot engines using unstructured documents”. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016. p. 516-525.

[5] Singh, J., Joesph, M. H., & Jabbar, K. B. A., “Rule-based chabot for student enquiries”. In: Journal of Physics: Conference Series. IOP Publishing, 2019. p. 012060.

[6] Zamora, J., “Rise of the chatbots: Finding a place for artificial intelligence in India and US”. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces Companion. 2017. p. 109-112.

[7] Sheikh, S. A., Tiwari, V., & Singhal, S., “Generative model chatbot for Human Resource using Deep Learning”. In: 2019 International Conference on Data Science and Engineering (ICDSE). IEEE, 2019. p. 126-132.

[8] Wang, Z., Wang, Z., Long, Y., Wang, J., Xu, Z., & Wang, B., “Enhancing generative conversational service agents with dialog history and external knowledge”. Computer Speech & Language, 2019, vol. 54, p. 71-85.

[9] Zhang, J., Huang, H., & Gui, G., “A Chatbot Design Method Using Combined Model for Business Promotion”. In: International Conference in Communications, Signal Processing, and Systems. Springer, Singapore, 2018. p. 1133-1140.

[10] Amudha, S., & Shanthi, I. E., “Phrase Based Information Retrieval Analysis in Various Search Engines Using Machine Learning Algorithms”. In: Data Management, Analytics and Innovation. Springer, Singapore, 2020. p. 281-293.

[11] Koka, R. S., “Automatic Keyword Detection for Text Summarization”. PhD diss., 2019.

[12] Hulth, A., & Megyesi, B. B., “A study on automatically extracted keywords in text categorization”. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006. p. 537-544.

[13] Fernandez, R. R., & Uy, C., “Keywords on Online Video-ads Marketing Campaign: A Sentiment Analysis”. Review of Integrative Business and Economics Research, 2020, vol. 9, p. 99-110.

[14] Siddiqi, S., & Sharan, A., Keyword and keyphrase extraction techniques: a literature review. International Journal of Computer Applications, 2015, vol. 109, no 2.

[15] Havrlant, L., & Kreinovich, V., “A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation)”. International Journal of General Systems, 2017, vol. 46, no 1, p. 27-36.

[16] Paukkeri M, Honkela T., “Likey: Unsupervised language independent keyphrase extraction”. In: Proceedings of the 5th international workshop on semantic evaluation, Uppsala, Sweden, 2010. p. 162-165.

[17] Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A., “YAKE! collection-independent automatic keyword extractor”. In: European Conference on Information Retrieval. Springer, Cham, 2018. p. 806-810.

[18] Nara, N., Sharma, P., & Kumar, P., “Page Rank Algorithm: big data analytic”. 2017.

[19] Mihalcea, R., & Tarau, P., “Textrank: Bringing order into text”. In: Proceedings of the 2004 conference on empirical methods in natural language processing. 2004. p. 404-411.

[20] Wan, X., & Xiao, J., “Single Document Keyphrase Extraction Using Neighborhood Knowledge”. In: AAAI. 2008. p. 855-860.

[21] Liu, Z., Huang, W., Zheng, Y., & Sun, M., “Automatic keyphrase extraction via topic decomposition”. In: Proceedings of the 2010 conference on empirical methods in natural language processing. Association for Computational Linguistics, 2010. p. 366-376.

[22] Mikolov, T., Chen, K., Corrado, G., & Dean, J., “Efficient estimation of word representations in vector space”. arXiv preprint arXiv:1301.3781, 2013.

[23] Hayati, H., Chanaa, A., Idrissi, M. K., & Bennani, S., “Doc2Vec &Naïve Bayes: Learners’ Cognitive Presence Assessment through Asynchronous Online Discussion TQ Transcripts”. International Journal of Emerging Technologies in Learning (iJET), 2019, vol. 14, no 08, p. 70-81.

[24] Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S., “Skip-thought vectors”. In Advances in neural information processing systems, 2015, pp. 3294-3302.

[25] Lau, J. H., & Baldwin, T., “An empirical evaluation of doc2vec with practical insights into document embedding generation”. arXiv preprint arXiv:1607.05368, 2016.

[26] Bennani-Smires, K., Musat, C., Hossmann, A., Baeriswyl, M., & Jaggi, M.,“Simple unsupervised keyphrase extraction using sentence embeddings. arXiv preprint arXiv:1801.04470, 2018.

[27] Sarosa, M., Junus, M., Hoesny, M. U., Sari, Z., & Fatnuriyah, M., “Classification Technique of Interviewer-Bot Result using Naïve Bayes and Phrase Reinforcement Algorithms”. International Journal of Emerging Technologies in Learning (iJET), 2018, vol. 13, no 02, p. 33-47.

[28] Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G., “Kea: Practical automated keyphrase extraction”. In: Design and Usability of Digital Libraries: Case Studies in the Asia Pacific. IGI global, 2005. p. 129-152.

[29] Nguyen, T. D., & Luong, M. T., “WINGNUS: Keyphrase extraction utilizing document logical structure”. In: Proceedings of the 5th international workshop on semantic evaluation. Association for Computational Linguistics, 2010. p. 166-169.

[30] G. Hinton et al., “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups”, IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97, 2012. Available: 10.1109/msp.2012.2205597.

[31] Amodei, D., Ananthanarayanan, S., Anubhai, R, et al., “Deep speech 2: End-to-end speech recognition in english and mandarin”. In: International conference on machine learning. 2016. p. 173-182.

[32] Lin, B., Zagalsky, A., Storey, M. A., & Serebrenik, A.. “Why developers are slacking off: Understanding how software teams use slack”. In : Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion. 2016. p. 333-336.

An Assessment of Organizational Capabilities for ERP Implementation in SMEs: A Governance Model for IT Success using a Resource-based Approach

Dr. Houcine Chatti[1], Dr. Evan Asfoura[2]
Dar Al Uloom University
Riyadh, Saudi Arabia

Dr. Gamal Kassem[3]
German University
Cairo, Egypt

***Abstract*—One of the most coveted technological innovations is the increasing use of Integrated Management Software (ERP) since the early 1990s. ERP is considered a powerful reengineering tool that profoundly transforms a company's business processes and changes the way to conduct reengineering projects and implement new software. The significant number of failures as reported in the literature on ERP emphasizes the fact that some companies may not realize the expected benefits. This result becomes particularly significant in the case of SMEs which have their own contingencies in addition to the scarcity of resources which may, in turn, lead to the failure of ERP implementation. This leads initially to ask, in one hand, about the variables of success of this innovation i.e. the determinants of the techno-organizational innovation; and on the other hand, about the existence of a model of dependences analysis between these determinants and their success as perceived by the management. The current empirical research is carried out in 92 companies having adopted a whole or a part of their IS with an ERP system. Having analyzed the data collected via a questionnaire, and applying the method of structural equation (MSE), results prove the existence of one "general fit" between the data and the supposed relations of causality.**

Keywords—ERP success; systemic approach; quantitative study; structural equation modelling

I. Introduction

ERP is considered a powerful reengineering tool that profoundly transforms a company's business processes and changes the way to conduct reengineering projects and implement new software. Many large companies are attracted by this IT because of their desire to evolve their architecture, but also the processes supporting the business system [1]. The 1st generation ERP was thus designed to process management information in real time and networked within the same organization. It helped streamline and integrate business processes and information flows, creating synergies between the company's resources. The 2nd generation system now offers a platform for reengineering and internal integration of business processes. It enables new models of inter-organizational integration by industry to be envisaged, as is the case in highly competitive sectors in which SMEs operate. The deployment of this IT is thus part of a context of governance that has been the subject of very little work.

Companies that choose an ERP face a double challenge: first, to control their choice [2], and second, to implement and use it with certain guarantees of success. In both cases, their managers must decide with knowledge of the risks of failure incurred [3], while ensuring, as Davenport [4] postulates, that this technology is thought out and used to produce real added value for the company. Thus, the interest in the failure, or conversely the success, of an ERP is justified in research work [5], but also by the specialized press, by the fact that the expected returns on this investment are difficult to quantify in terms of net results in the short or medium term, i.e. immediate direct benefits [6].

Investment in this enterprise software package is now more a standard for integrating management systems and as a strategic lever for intra versus inter-organizational collaboration [7]. Indeed, this standard makes it possible to coordinate operations in business networks by facilitating long-term relationships. It facilitates the integration of the management systems of SMEs belonging to the same industry or operating in the same market [8]. Marketplaces by industry are thus intended to support this collaboration by emphasizing the importance of being able to interconnect different management systems, whether CRM, B2B or SCM, to the company's ERP system.

This is particularly true for SMEs that have their own contingencies and limited resources, making them vulnerable to the failure of an ERP implementation. This observation leads us to assign a double objective to this work. Firstly, to identify the organizational and technological capabilities that are useful for accepting a technological innovation such as ERP and facilitating its success; and secondly, to propose a model for analyzing the effects of these capabilities on ERP success, in the case of SMEs. These objectives are therefore part of a general questioning on the respective and simultaneous influence (1) of specific and critical organizational capacities or resources (such as the capacities of the management, the IS and the software package, the implementation project,...); and (2) of capacities to innovate and implement this technological solution (such as the capacities to accept and integrate it in the management practices), on the performance of the company. This issue of evaluating the causes of an ERP's success, and the questions that emerge from this paper, firstly to identify the determinants of success, and secondly to formulate an original model for predicting success, are part of a general ERP governance research context [9].

II. Evaluating ERP Success

A. The Theoretical Context of the Evaluation

Evaluating the success of an ERP leads to dealing, from a theoretical point of view, with its governance, and finds its justification in the numerous studies [10] which aim to protect organizations from the risks of failure an implementation of this IT. In this sense, it is about producing stable, conceptually invariant knowledge and models that support proactive investment decisions. The evaluation thus responds to the basis of the model of a "logical platform" for developing the governance activities of an IT, formulated by Schwars and Hirscheim [11]. It concerns, according to Sambamurthy and Zmud [12], “contingency forces” (e.g. management and strengthening of organizational and technological capacities). It is these strengths that allow organizations to organize the governance mode of an ERP, by identifying, on the one hand, the critical possibilities or capacities that this technology offers (eg resources and creation of assets), and by designing, on the other hand, a management model (ie a relational versus integrated business architecture) capable of supporting its use and consistent with the strategic vision.

Two theoretical approaches, independent but complementary, are mobilized to identify the contingency forces which will be declined, in this work, in terms of resources or organizational and technological capacities in relation to the problem posed by the evaluation of the perceived success of an ERP. The first one borrows from the field of work on IT strategic planning, the hypothesis of the existence of determinants or key factors of success of an IT project. It is therefore a question of identifying certain resources, availabilities or organizational capacities that must be present when integrating the management system with an ERP. The second takes from the field of work on the diffusion of innovation the hypothesis that the process of adopting an IT depends on its intrinsic characteristics, which can also influence its success. It is therefore a question of identifying properties of this management process that transforms the existing organization as a result of the introduction of new technological capabilities.

To sum up, each of these two approaches contributes, with its own deterministic approach, to questioning the causes of the success of ERPs and the conditions of their governance, placing greater emphasis, as Barua and Mukhopadhyay [13] mention, on the behavioral and organizational capacities of organizations, whether economic or financial. The confrontation of these two approaches supports the deterministic research framework that has provided a large body of work over the last decade [6][14]. Much of this work has mobilized "behaviourist" models for evaluating IT success, very often derived from Davis [15] Technology Acceptance Model (TAM), or DeLone and McLean's [16] Adoption Model. The postulated complementarity of these approaches is in line with the theoretical format of these reference models and suggests, below, a justification of the theoretical choices and a generic approach for investigating research concepts based on a resource-based view of evaluating the success of ERPs.

B. The Theoretical Framework of Research

The theoretical perspectives mentioned are supported, initially, by the choice of strategic alignment models and organizational innovation, whose foundations, organizational "fit" and "adoption" of ERP, are established by multiple contingencies (e.g. managerial, organizational, strategic and technological). Each of these models supports an approach to evaluating ERP success - studies on "fit" seek to find interdependencies between various contingencies or capabilities that explain the success of this IT; those on "adoption" seek to identify the optimal conditions or capabilities for its adaptation to individuals and the organization, which also explain its success. The complementarity of the contributions of these two models is taken up by numerous works dealing with a resource-based approach (RBA) - this complementarity is addressed by the formulation of a systemic process of capacity transformation. The study of this process requires, in a second step, to formalize the structure of a theoretical working meta-model.

The Strategic IT Alignment Model, postulated by the founding model of Henderson and Venkatraman [17], is certainly the most appropriate model in the field of "ERP strategy" because it provides a formal structure for analyzing the key success factors. This structure, based on a study of the multiple contingencies of IT governance, is particularly well suited to work dealing with ERP performance, its direct or induced benefits, and the identification of its CSFs [18]. This model conceptually supports the involvement of the resources (managerial and technological) that an organization must have at its disposal to plan the development of an IT. The Strategic IT Alignment Model, which has been at the origin of numerous studies aimed at determining the impact of IT on organizational performance, is based in particular on the concepts of strategic integration and functional integration [19]. Strategic integration suggests an alignment between the external and internal business environment, while functional integration suggests the synergy between business processes and the IT used. The implementation of an ERP, leading to the integration of the organization's IT applications, limits here the theoretical investigation to functional integration alone. Only this integration makes it possible to apprehend the managerial, organizational and technological capacities required to properly align business processes (management model) with technical processes (integrated architecture model).

The theoretical framework for the diffusion of innovation, postulated by the founding model of Rogers [20], is also the one that has become established in the field of research on organizational change with IT, because it makes explicit the conditions for acceptance/adoption of an organizational innovation, and that of the management or resource allocation process that improves the contribution of ITs to the organizational performance of the firm. The original innovation adoption model, which has been the source of a great deal of work on IT, such as EDI or ERPs [21], emphasizes the perceived characteristics of an IT to explain the probability and speed of diffusion of innovation within the social system. Perceived characteristics play a fundamental role in the persuasion phase, during which the decision unit assesses the appropriateness of adopting or not adopting the technological innovation. In the case of ERPs, the work highlights the advantages, complexity and technical compatibility in a development project. This research retains only the "relative advantage" component as a characteristic attribute of an ERP innovation. This choice is explained, firstly, by the desire to simplify the ERP adoption model and, secondly, by the fact that the results of previous studies that have dealt with the adoption of this organizational innovation have concluded that only the "benefit provided" was a determinant of adoption [21] [10]. Thus, the main benefits identified were technical, operational, strategic, and organizational [22]. The technical benefits are primarily related to the capabilities of the IT infrastructure [23]. IT infrastructure is the basis and support for computer applications and, consequently, the prerequisite and determinant of any type of benefit provided by an IS [24]. This benefit (measured in terms of improving the capabilities of the IT infrastructure) would condition, in other words, the other benefits. It must, therefore, be considered as the one that is at the origin of the choice of an ERP innovation, and that leads to its adoption. The perceived characteristics of the innovation are therefore introduced into the modeling of ERP success in terms of "improving the capabilities of the IT infrastructure".

To summarize, the links between these two theoretical frameworks, supporting the premises of this work, find their justification in the work carried out by the RBV school "from a resource-based view of IT success evaluation" [25]. This posits, in a governance meta-model involving strategic technological and organizational resources, value-creating interactions. In this sense, it constitutes a conceptual aid for the analysis of IT adoption by considering the simultaneous effects of innovation and the integration of IT in an intermediate process of capacity transformation, as suggested by McKeen et al., [26]. This schema helps to delineate theoretical choices and interactions, and provides a revised framework for evaluating the success of IT. It helps to apprehend and justify the conceptual components in relation to (1) the functionalities of the IT studied, (2) the structural characteristics of the process of its adoption and its applications during its integration into the company's IS, and (3) the methods for evaluating the performance of the IT studied [27] [10]. Nevertheless, this research work will be limited to validating only the relationship between two levels of analysis of the theoretical construction: (1) the determinants of ERP success (technological and organizational); and (2) the success of ERP implementation (via the results induced by its use).

The research model (Fig. 1) establishes the relationships between (1) the key success factors of ERP implementation (prerequisites/precursors); and, (2) the result factors reflecting the benefits of the implementation (performance/success).

The existence of this relationship, established by previous work, is validated by the results of the exploratory study conducted by M.Abdel-hak et al., [27]. The interaction identified during this study between these two factors, which legitimizes a qualitative approach to the phenomenon of ERP adoption, which could for example be proposed for work on change management, i.e. in-situ and during implementation, is impossible to analyze in a context of ad hoc research limited to capturing the phenomenon at a given time. To be generalized, this model requires an empirical study conducted with a representative sample of SMEs, and a test of the data through hypothetico-deductive research. This study leads us to briefly recall, in this paper, the concepts and variables used for each of the two levels of this research model.

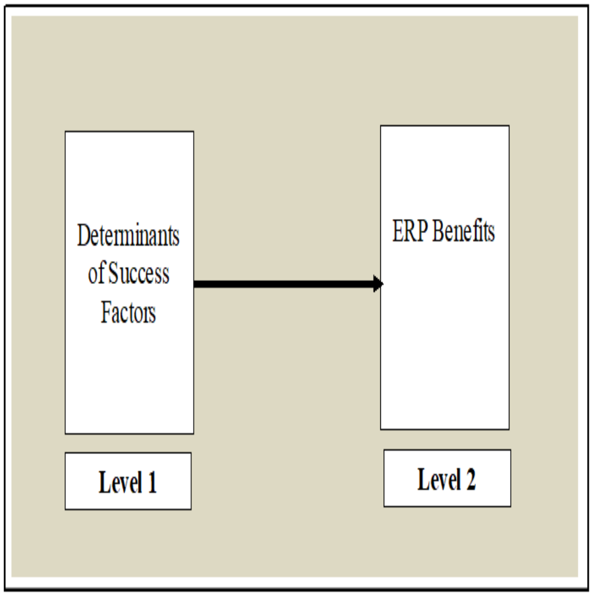


Fig. 1. General Structure of the Research Model.

III. The Conceptual Model of Research

A. *Benefits of Implementing an ERP System*

The results of the scoping study [27] indicate several potential benefits. Only those that appeared to be the most discriminating and that fall within the scope of the evaluation of ERP success. These intangible benefits relate to information literacy and improved skills utilization), and symbolize the improvement of individual capacities to do business in SMEs.

Information literacy is an important concept of ERP success and is the main benefit provided [28]. Previous work shows that one of the most important reasons why organizations adopt an ERP is its ability to facilitate the exchange of information, to solve information fragmentation problems [4], and to satisfy the information needs of all business units [29]. This concept is reflected in the results of the exploratory analysis in terms of availability (BEINFD), reliability (BEINFF), security (BEINFS) and real-time access (BEINFA) to information. These terms refer to variables to be explained in the model, to describe the improvement of the management and control of the information that circulates via the ERP. The comments collected, compared with the results of previous work, also show the organization's ability to control information, as a source of other induced benefits, such as improved decision-making, control, financial performance, productivity, resource and production management.

Improving skills utilization is an equally important concept of ERP success that should also be considered a significant benefit [23]. This concept is captured by the exploratory analysis in terms of the capabilities provided to employees to develop new and more appropriate management and organizational skills (BECPMT), as well as to promote better use of certain management tools and techniques (BECPTG). These terms refer to as many variables to be explained in the model measuring an improvement in the management of human resources, a source of wealth creation, but also a source of significant expenditure. This is why the optimal exploitation of employees' potential remains a permanent concern for companies. The results of previous work show that an ERP offers employees the opportunity to detach themselves from routine and repetitive tasks, and to concentrate on analytical tasks with more added value. The new business practices and the multiple functionalities offered by this IT also provide employees with the tools they need to better master procedures and develop their skills.

The success of an ERP implementation is evaluated, to summarize, by two conceptually independent levels of apprehension. The first, that of "information control" is the subject of four variables to be explained, measuring the efficiency of the management system impacted by information control. The second, that of "use of skills" is the subject of two variables to be explained, measuring the transformation and improvement induced in skills by the "fit" between work and IT.

B. *Determinants or Organizational Capacities for Success*

The results of the scoping study point to several key success factors, the most significant of which are the quality of the project team, the definition of the project team's mission, user training, user acceptance, commitment from senior management, business process re-engineering and selection of the ERP provider [27].

The quality of the project management team is a key concept in the success of an ERP implementation [30]. Previous work shows that a quality project team is a well-balanced, experienced team composed of qualified people [31], trained and led by a confirmed project manager. This concept is reflected in the results of the exploratory analysis in terms of the composition (DSEPCE) and competence (DSEPCQ) of the team.

The definition of the project management team's mission is a concept frequently cited as a determinant of success in ERP projects [32]. This concept is here unidimensional and is measured (DSEPME), reflecting the quality of ERP implementation project planning. It should be remembered that, in terms of ERP, the plan specifies the roles of the stakeholders and specifies the field of investigation, as well as the objectives of the project at these different stages.

User acceptance is also a concept frequently cited as a determinant of the success of ERP projects [33] [34]. This concept is here unidimensional and is measured (DSACCP), reflecting the fact that if users do not initially accept the system wholeheartedly, then senior management imposes constraints. Users are then forced to cooperate and deal with the system.

The lack of user training refers to a concept frequently cited as a cause of problems in ERP implementations [32]. This concept is one-dimensional and is measured (DSFORM), reflecting the fact that user participation in training sessions is necessary to operate the software package effectively. It should be remembered that, in ERP, training produces knowledge for all the functionalities and modules to be implemented.

The commitment of senior management refers to a concept that has been well identified, through work on IS planning, as a determinant of the success of IT, and therefore ERP, acceptance [32]. This concept is unidimensional and is subject to measurement (DSDGEN) with specific attributes that are important, primordial, strategic, philosophical, stimulating and driving. These terms are often used in the words of ERP project stakeholders to describe the role and commitment of senior management in the implementation process.

The Business Process Review refers to a generic concept (the BPR) that is always mentioned as an explicit condition for successful IS integration and e makes ERP implementation a success [35]. This concept is one-dimensional and is measured (DSRPA), which shows that the objective of this reconfiguration is generally to adapt the ERP to the solution desired by the customer, adapt the company's processes to the software package's standards, and improve business process performance.

The selection of the ERP solution refers to a concept, globally evaluating the quality and appropriation of the tender documents, which strongly conditions the success of ERP projects [32]. This concept is rendered by the results of the exploratory analysis in terms of the selection of the appropriate software package (DSFRCH) and the supplier/company relationship (DSFRRL). These terms are the subject of two different explanatory variables that measure: for the first, the fact that the ERP meets the organization's needs and the coverage of its processes, on which the success of its implementation and use will depend. The exploratory analysis shows, in this sense, that the majority of companies proceed, at the beginning of the project, to the elaboration of specifications based on an analysis of the needs and the offer on the market of ERP software packages. For the second, the fact that the supplier/enterprise relationship is vital to the success of the project [36]. The exploratory analysis also shows that the establishment of a trusting and serious relationship between the buyer and the ERP supplier is recommended for the successful conduct of the project.

IV. The Model and Research Hypothesis

The conceptual model (Fig. 2) supports, via a hypothetical-deductive approach, a systemic analysis framework designed to test the simultaneous influence of critical success factors (organizational capabilities) on the benefits of ERP implementation (business/managerial capabilities). The proposed deterministic framework, justified by previous exploratory work [27], assumes the existence of a network of direct causalities between endogenous and exogenous variables at both conceptual levels.

This empirical approach, conducted among SMEs that have implemented an ERP in recent years, postulates the existence of direct influential relationships between the variables of the system under study. This empirical study is thus intended to test a single general research hypothesis, which formulates the "fit" of the structure of the linear relations (R1) between the variables of the control model and the data. This "fit" is intended to identify an optimal model, from the only significant pathways between the endogenous and exogenous variables of the two conceptual levels (1 and 2).

The general hypothesis supporting this fact is limited, within the framework of this exploratory work, to stating these direct and indirect influences:

HG: The benefits of an ERP system (evaluated by the Level 2 variables) are directly influenced (R1) by organizational capacities (evaluated by the Level 1 variables).

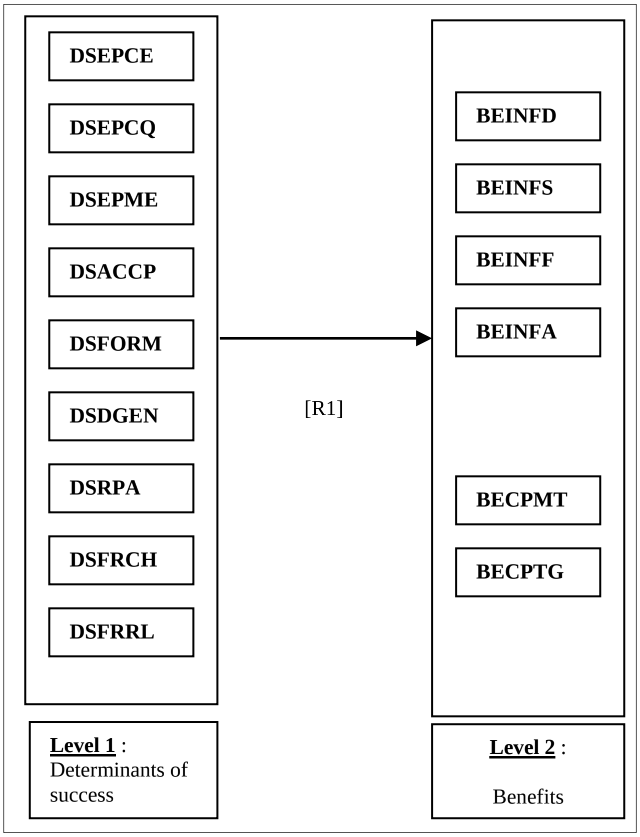


Fig. 2. The Conceptual Model.

V. The Methodological Framework of the Research

This study is carried out in the context of Saudi SMEs, which have opted for an ERP solution over the last decade. All the companies in the sample, 92 in all, are in a post-implementation phase of this IT at the time of the study, and are able to evaluate, ex-post, the conditions of implementation and the benefits of its use.

The practice evaluation questionnaire was developed based on the exploratory study [27] and a review of existing metrics in the literature in order to adapt them to the measurement of the research variables. This report presents two parts relating to the two components of the conceptual model. For each of these components, the concepts, variables and constructs are stated. Each construct is the subject of an item coding and a 5-point Lickert scale.

The questionnaire was administered to people who were both active members of the ERP project team and members of the general management, with a certain preference for the directors of information systems. Targeted individuals (such as CEOs, production, procurement, or sales managers) were expected to be involved in all phases of the project.

The treatment of the research problem is approached using a structured methodology using conjugation (justified by the low number of cases), and a sequence of 1st generation descriptive methods (such as principal component analysis (PCA) and the reliability test of constructs) and 2nd generation explanatory methods (such as structural equations and dependency analysis) with SPSS and AMOS.

VI. The Results of the Research

The results of the treatments carried out concern, for the variables of the model, the descriptive analysis intended to check the validity (convergent and discriminant) and the reliability of the constructs (Cronbach's "alpha" coefficients calculated for the items in the questionnaire. The results of the treatments for the HG test concern the study of the significance of the general fit of the model to the data and the study of the direct and indirect pathways between variables of the model.

TABLE I. The Results of the Reliability Test: (α) Cronbach

	Factors	**Items**	**Alpha (α)**
Level 1 (Determinants of success)	DSEPCE	Dsepce1,Dsepce2,Dsepce3	0,7023
	DSEPCQ	Dsepcq1,Dsepcq2,Dsepcq3,Dsepcq4	0,6979
	DSEPME	Dsepme1,Dsepme2,Dsepme3	0,7735
	DSACCPF1	Dsaccp2, Dsaccp3, Dsaccp4	0,9070
	DSACCPF2	Dsaccp5,Dsaccp6,Dsaccp7	0,8227
	DSFORM	Dsform1,Dsform2,Dsform3,Dsform4	0,6943
	DSDGENF1	Dsdgen1,Dsdgen2,Dsdgen5	0,8132
	DSDGENF2	Dsdgen3, Dsdgen4, Dsdgen5	0,6917
	DSRPA	Dsrpa1,Dsrpa2,Dsrpa3,Dsrpa4	0,8506
	DSFRCHF1	Dsfrch1,Dsfrch2,Dsfrch5,Dsfrch6	0,8558
	DSFRCHF2	Dsfrch3, Dsfrch4	0,5281
	DSFRRL	Dsfrrl1,Dsfrrl2,Dsfrrl3,Dsfrrl4	0,8624
Level 2 (Benefits)	BEINFD	Beinfd1, Beinfd2, Beinfd3	0,7468
	BEINFS	Beinfs1, Beinfs2, Beinfs3	0,7355
	BEINFF	Beinff1, Beinff2, Beinff3	0,7880
	BEINFA	Beinfa1,Beinfa2,Beinfa4	0,8140
	BECPMT	Becpmt1, Becpmt3, Becpmt4, Becpmt5, Becpmt6	0,8450
	BECPTG	Becptg1, Becptg2, Becptg3	0,8436

TABLE II. CONVERGENT VALIDITY OF ONE-DIMENSIONAL CONSTRUCTS

DSFORM	Items	1
	Dsform1	,739
	Dsform2	,814
	Dsform3	,694
	Dsform4	,655
	Eigenvalues	2,119
	Explained variance	52,966 %
	Cumulative variance	52,966 %
DSRPA	Items	1
	Dsrpa1	,761
	Dsrpa2	,860
	Dsrpa3	,840
	Dsrpa4	,879
	Eigenvalues	2,795
	Explained variance	69,883%
	Cumulative variance	69,883%
DSEPCE	Items	1
	Dsepce1	,684
	Dsepce2	,794
	Dsepce3	,905
	Eigenvalues	1,918
	Explained variance	63,918 %
	Cumulative variance	63,918 %
DSEPME	Items	1
	Dsepme1	,831
	Dsepme2	,782
	Dsepme3	,880
	Eigenvalues	2,076
	Explained variance	69,202 %
	Cumulative variance	69,202 %
DSEPCQ	Items	1
	dsepcq1	,782
	dsepcq2	,588
	dsepcq3	,785
	dsepcq4	,746
	Eigenvalues	2,131
	Explained variance	53,271%
	Cumulative variance	53,271%

A. Content Validity of the Research Variables

1) Variable content validity - organizational capabilities (Level 1): Each construct measuring a variable or determinant of ERP success at the first conceptual level is subject to a validity test with a principal component analysis (PCA) and a reliability test.

The principal component analysis (PCA) of the attributes of the "user training" (DSFORM), "business process reengineering" (DSRPA) and "project team mission" (DSEPME) constructs shows the existence of a one-dimensional factorial structure. The dimension extracted by the principal component analysis (PCA) has an eigenvalue equal to 2.076 and restores an explained variance higher than 69%. The factorial contributions of the measurement items are greater than 0.654. The reliability of this construct is confirmed. (Table I)

The quality of the project team is assessed by two variables measuring the composition of the project team (DSEPCE) and the capacity of the project team (DSEPCQ). principal component analysis (PCA) of the attributes of the first construct measuring "the composition of the project team" and "the capacity of the project team" verify their unidimensionality. The extracted factor has an eigenvalue greater than 1.918 and returns an explained variance greater than 53.271%. The factorial contributions of the three measurement items are greater than 0.60 (Table II). The reliability of these constructs is confirmed (Table I).

The principal component analysis (PCA) of the attributes of the constructs "user acceptance" (DSACCP) and "management commitment" (DSDGEN) shows the presence of a factor structure with two independent components. These extracted factors jointly restore an explained variance greater than 74.042% and each has an eigenvalue greater than 1. For the rest of the study, the "Dsaccpf2" factor is retained as an effective measure of user acceptance, while the "Dsaccpf1" factor is retained as a measure of ease of use. Similarly, the "dsdgenf1" factor was chosen as an effective measure of the involvement of the general management, while the "dsdgenf2" factor was chosen as a measure of the general management's intervention (Table III). The reliability of these two constructs is confirmed (Table I).

The selection of the ERP solution is measured by two supposedly independent variables, "the selection of the ERP software package (DSFRCH)" and "the quality of the company/supplier relationship (DSFRRL)".

TABLE III. CONVERGENT VALIDITY OF TWO-COMPONENT CONSTRUCTIONS

DSACCP	Items		
		Dsaccpf1	Dsaccpf2
	Dsaccp1		,517
	Dsaccp2	,438	
	Dsaccp3	,852	
	Dsaccp4	,928	
	Dsaccp5	,895	,782
	Dsaccp6		,836
	Dsaccp7		,849
	Eigenvalues	3,850	1,333
	Explained variance	55%	19,042%
	Cumulative variance	55%	74,042%.
DSDGEN	Items		
		Dsdgenf1	Dsdgenf2
	Dsdgen1	,868	
	Dsdgen2	,906	
	Dsdgen3		,846
	Dsdgen4		,800
	Dsdgen5	,785	,784
	Eigenvalues	2,090	1,809
	Explained variance	41,797%.	36,189%
	Cumulative variance	41,797%	77,986%

The Principal Component Analysis (PCA) of the attributes of the first construct "selection of the ERP software package" shows the presence of a two-dimensional factor structure. The items have high contributions on the two extracted factors (> 0.80). The first factor "Dsfrchf1", obtained from attributes 1, 2, 5 and 6, is retained as an effective measure of the project requirement, with a high eigenvalue of 2.823, restores an explained variance of 47.054%. The second factor "Dsfrchf2", obtained from items 3 and 4, is retained as a measure of the adaptation of the offer, with an eigenvalue of 1.459, restores an explained variance of 24.320%. The cumulative variance of these two factors is 71.374% (Table IV). The reliability of these two constructs is confirmed (Table I).

The principal component analysis (PCA) of the attributes of the second construct measuring the "quality of the company/supplier relationship" shows the presence of a one-dimensional factorial structure. The factorial contributions of these measurement items are greater than 0.769. The extracted dimension has a high eigenvalue of 2.853 and returns 71.313 % of explained variance.

2) Content validity of the variables to be explained - managerial skills (Level 2): The second conceptual level refers to two multidimensional concepts related to "information literacy" and "skills use".

The concept of "information literacy" is assessed by the variables BEINFD, BEINFS, BEINFF, BEINFA. The four constructs are measured by twelve attributes. The principal component analysis (PCA) of these attributes gives four factors. The factorial contributions, higher than 0.558, on the four factors returned (Table V) are all significant.

The concept of "best use of skills" is assessed through the variables BECPMT, BECPTG. Both constructs are measured by eight attributes. The principal component analysis (PCA) of these attributes restores two factors. The factorial contributions, higher than 0.602, on the two restituted factors (Table VI) are all important.

B. Testing the General Hypothesis HG

The purpose of this test is to validate the general structure of the control model. It is carried out by studying, firstly, the "fit" of the general model and, secondly, the paths in an optimal model rejecting the path coefficients between variables to be explained and non-significant explanatory variables (p>5%).

The evaluation of the quality of the fit of the model to the data consists in estimating the quality of the fit of the theoretical model to the empirical data. This step (in particular the analysis of the precision indices of the global model) uses a combination of indices with different characteristics (absolute, incremental and parsimony indices). The analysis of the values of these indices indicates acceptable results; the model is "over-identified" (the degree of freedom (ddl) is equal to 120 and therefore strictly positive) and the set of indices respects the limit values suggested by the literature on the structural equation method (Table VII).

The analysis of the quality of fit, with these indices (Table VI), shows an excellent fit (p=88.4%) of the proposed model to the data, as well as a very small impact of residuals on the model (5.6%). The general assumption regarding the overall structure of the model and in particular the presence of causal effects between the different explanatory and explanatory variables of the model is thus generally accepted. It is therefore possible to state that the benefits of an ERP system are directly influenced by the key success factors of the ERP implementation.

TABLE IV. CONVERGENT VALIDITY OF THE CONSTRUCTED "SELECTION OF THE ERP SOFTWARE PACKAGE"

	Items		
		Dsfrchf1	Dsfrchf2
DSFRCH	Dsfrch1 Dsfrch2 Dsfrch3 Dsfrch4 Dsfrch5 Dsfrch6	,821 ,859 ,780 ,863	,804 ,813
	Eigenvalues	2,823	1,459
	Explained variance	47,054%	24,320%
	Cumulative variance	47,054%	71,374%
DSFRRL	Items	1	
	Dsfrrl1 Dsfrrl2 Dsfrrl3 Dsfrrl4	,864 ,888 ,769 ,851	
	Eigenvalues	2,853	
	Explained variance	71,313 %	
	Cumulative variance	71,313 %	

TABLE V. DISCRIMINATORY VALIDITY OF THE CONCEPT OF INFORMATION LITERACY

Items				
	BEINFD	BEINFS	BEINFF	BEINFA
BEINFD1 BEINFD2 BEINFD3 BEINFS1 BEINFS3 BEINFS4 BEINFF1 BEINFF2 BEINFF3 BEINFA1 BEINFA2 BEINFA4	,828 ,840 ,705	,558 ,851 ,805	,754 ,616 ,884	,892 ,727 ,672

TABLE VI. DISCRIMINATORY VALIDITY OF THE BETTER USE OF SKILLS CONCEPT

Items		
	BECPMT	BECPTG
BECPMT1 BECPMT3 BECPMT4 BECPMT5 BECPMT6 BECPTG1 BECPTG2 BECPTG3	,644 ,875 ,602 ,762 ,805	,888 ,910 ,661

TABLE VII. PRECISION INDICES FOR THE GENERAL MODEL

	Precision indices	Calculated values	Recommended values
Absolute indices	Chi-Deux (χ^2)	96,4	
	P (χ^2= 96,4)	**88,40 %**	>10%
	ddl	120	
	χ^2/ ddl	**0,738**	< 5
	GFI	**0,898**	> 0,9
	AGFI	0,873	> 0,9
	RMR	**0,046**	< 0,1
	RMSEA	**0,000**	< 0,08
Incremental indices	NFI	0,883	> 0,9
	CFI	1,000	> 0,9
Parsimony indices	PNFI	0,618	
	PGFI	0,643	

The validation of HG also requires the verification of the significance of the dependency model parameters and the determination coefficients (R²) of the variables to be explained of the ERP implementation benefit.

Analysis of the regression coefficients shows that the direct dependency relationships are significant at the risk threshold p < 5%. These relationships reflect a high determinism of the influence of the explanatory variables on the ERP implementation result variables.

The determination coefficients (R²) of the dependent variables are high and significant (P< 1%). Indeed, 39.2% of the variance of the information availability variable "BEINFD", 70.2% of the variance of the information security variable "BEINFS", 57.3% of the variance of the information reliability variable "BEINFA", 64.2% of the variance of the information reliability variable "BEINFF", 75.3% of the variance of the variable "BECPMT" and 69.4% of the variance of the variable "BECPTG" are explained by the causal model HG.

Thus, the results relating to the "fit" of the model and the explained variance of the dependent variables validate the structure of the model tested. The optimal model that supports the HG hypothesis (representing the only significant dependency relationships identified at the significance level <5%) is established by the following relationship diagram (Fig. 3).

The pathway analysis performed for these only significant pathways (p < 5%) partially confirms the HG hypothesis. The pathway study reveals the main dependency relationships between the benefit variables of an ERP system (Level 2), and the key success factors (Level 1). The results of the pathways study carried out on the direct dependency relationships allow us to confirm globally the hypotheses of direct influences of organizational and technological capacities on managerial capacities for information management and improvement of the use of skills (Table VIII).

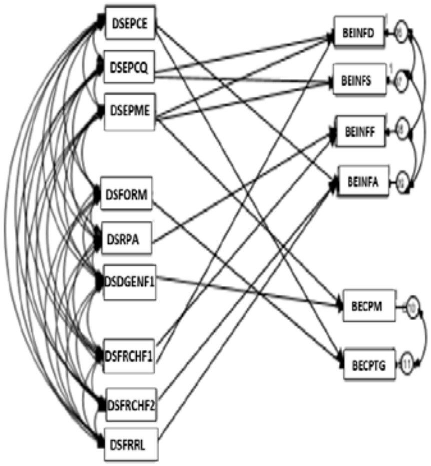


Fig. 3. The Optimal Control Model HG.

TABLE VIII. TESTING OF ADJACENT RESEARCH HYPOTHESES

Hypotheses
The composition of the project team DSEPCE has a positive impact on the four dimensions of information mastery (availability (BEINFD), reliability (BEINFF), security (BEINFS) and access (BEINFA)).
The composition of the DSEPCE project team has a positive impact on improving the use of skills (BECPMT).
The composition of the project team DSEPCE has a positive impact on improving the use of certain management tools and techniques (BECPTG).
The capacity of the project team (DSEPCQ) has no influence on the availability of information (BEINFD)
The composition of the project team (DSEPCE) as a positive impact on improving the use of skills.(BECPMT)
The mission of the project team (DSEPME) has a positive impact on information security and access to information ("BEINFS", "BEINFA").
The mission of the project team (DSEPME) has a positive impact on improving the use of skills (BECPMT).
Training (DSFORM) has a positive impact on the availability of information (BEINFD).
Training (DSFORM) has a direct positive impact on improving the use of management tools and techniques (BECPTG).
Business process reengineering (DSRPA) has a positive impact on the four dimensions of information mastery (BEINFD, BEINFS, BEINFA, BEINFF)
Business Process Re-engineering (DSRPA) has a positive impact on both dimensions of skills utilization improvement ("BECPMT", "BECPTG").
The requirement of the ERP project "DSFRCHF1" has a positive impact on the four dimensions of information mastery ("BEINFD", "BEINFS", "BEINFA", "BEINFF")
The requirement of the ERP project "DSFRCHF1" has a positive impact on both dimensions of skills utilization improvement ("BECPMT", "BECPTG").
The adaptation of the ERP offer (DSFRCHF2) has a positive impact on access to information (BEINFA).
The company/ERP supplier relationship (DSFRRL) has a positive impact on the access to information (EINFA).
The company/ERP supplier relationship (DSFRRL) has a positive impact on the new management and organizational skills (BECPMT).
The involvement of senior management in the management of resources (DSDGENF1) has a negative influence on access to information (BEINFA).
The intervention of the general management (DSDGENF2) has no influence on the ERP benefits.
User acceptance of (DSACCP) has no influence on the benefits of ERP implementation.

VII. Discussion of the Results

The results obtained from the path analysis of the optimal search model lead to different conclusions regarding the determinants of ERP success:

The "quality of the project team" is one of the most decisive factors for ERP success. Companies planning to implement an ERP system need to ensure that they have a balanced project team (a combination of both technical and managerial skills) with the knowledge, talent, and experience to handle the length and complexity of the task [37].

The positive impact of project team composition on information literacy is due to the dual skills (technical and managerial) of the project team. This mix of skills facilitates the understanding and technical formulation of users' information needs. This is a necessary step in the operation of parameterizing the ERP system and is important for improving information sharing [38] and availability [39].

The positive impact of the project team's composition on improving the use of skills also shows that when the team is balanced, the choices relating to the parameterization operation (the functionalities to be used), to the management of access to the system's applications, better meet the needs of users and better promote the use of talents and experience.

The definition of the "mission of the project team" is essential for the proper conduct of the ERP project. Specifying the role of each party involved in the project (at contract or specification level, even before the call for tenders is issued) makes it possible to establish the responsibilities and rights of each party [40]. The precision of the objectives of the ERP project also makes it possible to achieve the benefits expected from the implementation of the ERP system.

Business process re-engineering" is necessary to achieve (directly and/or indirectly) the projected benefits of implementing ERP systems [41]. The incompatibility of the characteristics of ERP software packages with the organization's business processes and the IT infrastructure in place may be the source of problems in implementing the ERP system. It is imperative for the success of ERP system implementations that the ERP implementation be preceded, therefore, by a review of the content of the functions, processes, tasks and IT infrastructure in place [39]. The objective of this reconfiguration is generally to respond to the solution desired by the client, to adapt processes to the standards of the ERP software package and/or to optimize the use of the ERP's functionalities.

The impact of reengineering on the use of new management tools and techniques is explained by the effect of "Best Practices" and the rich range of functionalities incorporated in ERPs. The business processes, incorporated in the ERP, are the result of long experience and benchmarking operations in different economic sectors. They embody the best business practices that can generate a competitive advantage for those who adopt it [42]. For many companies, complying with these rules and procedures translates into significant productivity gains [4].

The positive impact of reengineering on improving employee work control is due to the revision of the content of their functions and tasks. The standardization of company processes means that users are obliged to improve their old management and organizational practices.

The positive impact of re-engineering on the improvement of information control is due to the rationalization of the IT platform. This standardized platform allows a diverse use of information (e.g. summarized, aggregated, condensed data from several information sources within the company). Upgrading the IT infrastructure usually results in the replacement of transactional file systems or disparate databases with a single relational database system. These RDBMSs allow for better data management and manipulation.

User training" is a prerequisite for the success of the ERP project. The participation of users in training sessions is necessary for the efficient operation of the software package. These training actions must be conducted according to a clear program that is capable of meeting the expectations of the users. It is also recommended that training be carried out on all the functionalities and modules to be implemented for a better appreciation of the functional links generated by the integration. Lack of user training frequently appears to be responsible for problems in implementing ERP systems [43].

The positive impact of training on the availability of information is explained by the fact that the training provided on the various functionalities of the ERP system allows ERP users to be much more autonomous with regard to information. The training focuses on the possibilities offered by the new system to each user to have and use the information he needs, from his workstation.

The positive impact of training on improving the use of BECPTG management tools and techniques shows that training can benefit employees through the acquisition of certain management and organizational skills. The training allows employees to make better use of certain management tools and techniques, such as "dashboards, simulation techniques" incorporated into the system, to practice "international management", to "break away from routine tasks, focus on business aspects and move on to analytical tasks".

The "selection of an ERP solution" appropriate to the company's needs is a condition for the success of ERP projects [44]. KSA companies call on the services of their ERP suppliers to help them implement the system they have acquired. In this case, they are called upon not only to seek out the software package that best suits their needs, but also to ensure the choice of a supplier with the human resources necessary for the proper conduct of the project.

The precision of the "ERP project requirements" for the choice of a better ERP solution is an important factor for success. Before acquiring an ERP system, companies generally carry out a needs analysis and an analysis of the offer on the market for ERP software packages. Choosing the wrong software package can mean committing to an IT architecture and applications that do not fit the organization's objectives [45] or business processes and therefore limits the chances of ERP success.

The positive impact of training on improving the use of BECPTG management tools and techniques shows that training can benefit employees through the acquisition of certain management and organizational skills. The training allows employees to make better use of certain management tools and techniques, such as "dashboards, simulation techniques ..." incorporated into the system, to practice "international management", to "break away from routine tasks, focus on business aspects and move on to analytical tasks".

The "selection of an ERP solution" appropriate to the company's needs is a condition for the success of ERP projects [44]. Tunisian companies call on the services of their ERP suppliers to help them implement the system they have acquired. In this case, they are called upon not only to seek out the software package that best suits their needs, but also to ensure the choice of a supplier with the human resources necessary for the proper conduct of the project.

The precision of the "ERP project requirements" for the choice of a better ERP solution is an important factor for success. Before acquiring an ERP system, companies generally carry out a needs analysis and an analysis of the offer on the market for ERP software packages. Choosing the wrong software package can mean committing to an IT architecture and applications that do not fit the organization's objectives [45] or business processes and therefore limits the chances of ERP success.

VIII. Conclusion

The results of the analyses designed to test the partial hypotheses of dependence, confirm the existence of a general structure of the research model at two conceptual levels, and verify the general hypothesis in its generality. It should be noted that the general hypothesis is generally accepted and, consequently, that the benefits of using an ERP are influenced by the key success factors of the ERP implementation. Aside from the intervention of general management and user acceptance, which have no effect on the success of ERP implementation, the analysis has validated the benefits and the FCS identified during the exploratory study [27].

The spin-offs of this research work concern the development of companies. They are both beneficial for those who have already adopted an ERP system, but also for those who plan to acquire and implement such a tool. For the first category, a framework for evaluating the success of the ERP project is developed. This framework makes it possible to justify this investment by evaluating its benefits. For the second category, a coherent framework of ERP appropriation based on the experience of the first ones is formulated and to take advantage of it (based on the results of this work). Indeed, companies can expect to improve their performance and reap the benefits of an ERP implementation, knowing that achieving this objective is conditioned by a number of key success factors that determine the benefits of implementing an ERP. These key success factors are identified in this work in order to contribute to better planning and management of this type of project. These contributions provide the opportunity, among other things, to propose means of action (direct and indirect) relating to the determining factors of ERP success.

On the theoretical level, while helping to validate some theoretical and conceptual results of previous work, this work contributes to a better understanding of the success of ERP adoption in an emerging country, by proposing a model for evaluating ERP success, according to a systemic approach, thus resembling the necessary governance of ERPs on the strategic objectives of the SMEs studied. This modeling is one of the main contributions of the research because it mobilizes specific research variables that can enrich the modeling of this type of problem.

Acknowledgment

The researchers extend their thanks and gratitude to the Deanship of Graduate Studies and Scientific Research at Dar Al Uloom University for their support and funding of this study.

References

[1] Laudon, K. C. et Laudon, J. P., Management Information Systems: Managing the Digital Firm. New-York: Prentice Hall, 2018.

[2] EPICOR, 11 Criterions for Selecting the Best ERP System Replacement,http://www.epicor.com/pages/default.aspx, ,2009.

[3] Hakim, A. et Hakim, H., “A Practical Model on Controlling the ERP Implementation Risks”, Information Systems, Vol. 35, N° 2, April, 2010, pp. 204-214.

[4] Davenport, T. H., “Mission Critical”, Harvard Business School Press, Boston, Massachusetts, 2000.

[5] Shih-Ya, Y., “The Effect of Computer Self -Efficacy on Enterprise Resource Planning Usage”, Behavior & Information Technology, 2006, Vol. 25, No. 5.

[6] Chou, S. W. et Chang, Y.C., “The Implementation Factors That Influence The ERP (Enterprise Resource Planning) Benefits”, Decision Support Systems, Vol. 46, N° 1, December, 2008, pp. 149-157.

[7] Baile, S. et Trahand, J., “Les systèmes d’information inter-organisationnels – contributions et cadre de recherche”, Système d’Information & Management, Vol. 4, N° 2, 1999, pp. 2-19.

[8] Baile, S. et Mékadmi, S. , La conduite du changement dans un projet ERP: deux retours d’expériences, 11ème Congrès de l’AIM, Luxembourg , 2006 ,.

[9] Baile, S. , Le choix d’une forme de gouvernance organisationnelle des technologies de l’information et de leur succès – Prédictions avec les théories des Coûts de Transaction, du Structuralisme et des Ressources. Papier de Recherche du CERGAM, IAE Aix en Provence. (Soumis au 19ème Congrès AIMS’2010, Luxembourg), 31 pages.

[10] Chatti, H.,“Les déterminants du succès d’une implantation d’un ERP dans le contexte tunisien**”,** Thèse de Doctorat en Sciences de Gestion, CRG-IAE, Université de Toulouse 1, Juillet, . 2008, 540 pages.

[11] Schwars, A. et Hirscheim, R. , “An Extended Platform Logic Perspective of IT Governance: Managing Perceptions and Activities of IT, ” Strategic Information Systems, Vol. 12, 2003, pp. 129-166.

[12] Sambamurthy, V. et Zmud, R., “Arrangements for Information Technology Governance: A Theory of Multiple Contingencies, ” MIS Quarterly, Vol. 23, N° 2, 1999, pp. 261-290.

[13] Barua, A. et Mukhopadhyay, T., “Information Technology and Firm Performance: Past, Present and Future,” in Zmud, R.W., (Ed.), Framing the Domain of IT Management research: Projecting the Future Through the Past. Pinnaflex Educational Resources, 2000, Cincinnati, OH.

[14] Kerimoglu, O., Basoglu, N. et Daim, T., “Organizational Adoption of Information Technologies: Case of Enterprise Resource Planning Systems”, Journal of High Technology Management Research, Vol. 19, 2008, pp. 21–35.

[15] Davis, F.D. , “Perceived Usefulness, Perceived Ease of Use and User Acceptance Information Technologies, MIS Quarterly, Vol. 13, 1989, pp. 319–340.

[16] DeLone W.H. et McLean E.R. , “The DeLone and McLean Model of Information Systems Success: a Ten-Year Update”, Journal of Management Information System, Vol. 19 N° 4, 2003, pp. 9–30.

[17] Henderson, J. et Venkatraman, N., “Strategic Alignment: Leveraging information technology for transforming organizations”, IBM Systems Journal, Vol. 32, 1993, pp.4-16.

[18] Velcu, O. ,“Strategic Alignment of ERP Implementation Stages: An Empirical Investigation”, Information & Management, (ScienceDirect :Forthcoming), 2010,.

[19] Jouirou, N. et Kalika, M., “ l’alignement stratégique : déterminant de la performance - une étude empirique sur les PME ”, 9ème Congrès de l’AIM, Evry, 2004.

[20] Rogers, E.M., Diffusion of Innovations, New -York, Free Press (5th Ed), 2003.

[21] Ettien, F.A.K, L'impact individuel des ERP, un axe d'évaluation des changements induits. Thèse de Doctorat en Sciences de Gestion, CRM-IAE Toulouse, Université de Toulouse 1-Capitole, Février, 2007,.

[22] Pupion, P.C et Leroux E, “ Diffusion des ERP et comportements mimétiques ”, 15ème Conférence Internationale de Management Stratégique (AIMS), 2006, Annecy-Genève, Juin.

[23] Shang, S. et Seddon, P.B., A Comprehenive Framework For Classifying the Benefits of ERP Systems, Proceedings of the 2000 Americas Conference on IS, Long Beach California, August 10-08-2000, pp. 1005-1014.

[24] Kumar V et al, , “ERP Systems Implementation: Best Practices in Canadian Government Organizations”, Government Information Quarterly, N° 19, 2002, pp.147–172.

[25] Wade, M. et Hulland, J. ,“The Resource-Based View and Information Systems Research: Review, Extension and Suggestions for Future Research”, MIS Quarterly, Vol. 28, N°1, 2004, pp. 107–142.

[26] McKeen, J.D., Smith, H.A. et Parent, M., "Assessing the Value of IT: The Leverage Effect", Fourth European Conference on Information Technology, Cork, Ireland, June 21-13, 1997.

[27] M.S Abdel-Haq, H Chatti, E Asfoura,, Investigating the Success and the Advantages of Using ERP System in KSA Context , published by Engineering, Technology & Applied Science Resea, vol, 8 Issue 6, 2018.

[28] Andreas, I. et Nicolaou, S., “Firm Performance Effects in Relation to the Implementation and Use of Enterprise Resource Planning”, Journal of Information Systems, Vol. 18, N° 2, 2004, pp. 79-105.

[29] Davenport, T., “Putting the Enterprise into the Enterprise System”, Harvard Business Review, Vol. 76, N° 4, 1998, pp. 121-131.

[30] Nah, H. et Fui, F. , “Toward a Greater Understanding of End-User's Acceptance of ERP systems”, in Kosrow, M. (Ed.): Advanced Topics in Information Resources Management, Vol. 5, 2006, Idea Group Pub.

[31] Nah, F. et Delgado, S., “Critical Success Factors For Enterprise Resource Planning Implementation and Upgrade”, The Journal of Computer Information Systems, Vol. 46, N° 55, 2006, pp. 99-113.

[32] Ngai, E. et al., “Examining the Critical Success Factors in the Adoption of Enterprise Resource Planning”, Computers in Industry, doi:10.1016/j.compind., 2008.

[33] Netemeyer, R. G., Bearden, W. O. et Sharma, S., Scaling Procedures: Issues and Applications. New-York: Sage Publications, 2003.

[34] Seymour, L. et al., “End-Users Acceptance of Enterprise Resource Planning Systems: An Investigation of Antecedents”, Proceedings of the 6th Annual ISONE world Conference, April, 2007, Las Vegas, NV., available at http://www.isoneworld.org.

[35] Malhotra, R. et Temponi, C., “Critical Decisions for ERP Integration: Small Business Issues”, International Journal of Information Management, Vol. 30, N° 1, February, 2010, pp. 28-37.

[36] Raymond, L. et Uwizeyemungu., S., “A Profile of ERP Adoption in Manufacturing SMEs”, Journal of Enterpnse Information Management, Vol. 20, N° 4, 2004.

[37] Wor1ey, J. H., Chatha, K.A., Weston, R.H.., Aguirre, O. et Grabot. B., “Implementation and Optimisation of ERP Systems: a Better Integration of Processes, Roles, Knowledge and User Competencies”, Computers in Industry, Vol. 56, N° 6, 2005, pp 620-638.

[38] Singletary, L,“Empirical Study of Attributes and Perceived Benefits of Applications Integration for Enterprise Systems”. PhD Dissertation, Louisiana State University, 2003,.

[39] Ehie, I. et Madsen, M., “Identifying Critical Issues in Enterprise Resource Planning (ERP) Implementation”, Computers in Industry, Vol. 56, 2005, pp 545-557.

[40] Parr, A.N., Shanks, G. et Darke, P., “The Identification of Necessary Factors For Successful Implementation of ERP Systems”, in O.Ngwenyama, L.D.Introna, M.D.Myers et J.I. De Gross (Eds), New Information Technologies in Organisational Process. Boston: Kluwer Academic Publishers, 1999, pp. 99-119.

[41] Aloini, D., Dulmin, R. et Mininno,V., “Risk Management in ERP Project Introduction: Review of the Literature”, Information & Management, Vol. 44, N° 6, September, 2007, pp. 547-567.

[42] Sirigindi Subba Rao, , “Entreprise Resource Planning in ReengineeringBbusiness”, Business Process Management Journal, Bradford, 2000.

[43] Nelson K. et Somers T. M., “The Impact of Critical Success Factors across the Stages of ERP Implementations”, Proceedings of the 34th Hawaii International Conference on System Sciences, 2001.

[44] Wu, J.H.; Shin, S.S.; Heng, M.S.H., “A Methodology for ERP Misfit Analysis”; Information & Management, 44, 2007, pp 666–680.

[45] Robinson et Dilts,, “Or & ERP : a Match for the New Millennium ?”, ORMS Today, Vol.26, N°3, 1999, pp. 30-35.

On Developing High-Speed Heterogeneous and Composite ES Network through Multi-Master Interface

J Rajasekhar[1], Dr. JKR Sastry[2]
Koneru Lakshmaiah Education foundation Vaddeswaram

***Abstract*—These days, many heterogeneous and composite embedded systems contain many subnets developed using different bus-based protocols, such as I2C, CAN, USB, and RS485. There is always a requirement to Interface and interconnect the heterogeneous ES networks to achieve and establish a composite network. The ES networks developed using different protocols differ in many ways, considering the speed of communication, Arbitration, Synchronization, and Timing. Many solutions are being offered using heterogeneous embedded systems, especially in implementing automation systems, without addressing integration and proper interfacing. In this paper, a Multi-Master based interfacing of a CAN and I2C networking through Ethernet-based interfacing has been presented especially to find the optimum speeds at which the networks must be operated for different data packet sizes. It has been shown in the paper that it is quite efficient and effective when a data packet of size 40 bytes is driven using an I^{2}C speed of 5120 bits, Ethernet speed of 20480 bits, and CAN speed of 500 bits.**

Keywords—Embedded systems; embedded networks; hybridization of embedded networks; hybridizations through multi-master communication

I. Introduction

Many kinds of embedded networks are in existence and in use for implementing different types of Applications. The Most important Embedded System networking being in use includes the networks built using the communication systems that Include I2C, CAN, RS485, USB. But as the technologies are emerging, a necessity arises that require bridging the ES networks built around ES networking standards.

All ES networking standards differ in many ways: network termination, device identification, type and format of data packets, type of signals used, and communication speeds.

Hybridization of Embedded Networks can be achieved through different interconnecting types of wired networks that include I2C, CAN, USB, and RS48. Serial communication takes place among the hybridized networks. Hybridization can also be achieved through establishing wireless networks or through a combination of wireless and wired networks. The major issue in such networking is the management of communication speeds and data rates. May architectures can achieve the networks' hybridization, including single master integrations, multi-master integrations, a hardware-based bridge, multi-master integration, etc. The way a hybridized system works is dependent on the type of hybridization method used. As of date, hybridization through a Bridge device is proposed by [17].

Hybridization can be achieved through other methods that include Single master catering for the communication, using a Multi-master interface, and developing a Universal bus that caters to most of the ES-based communication Standards. In this paper, hybridization through Multi-master Interface has been presented.

A. Motivation

The interfacing of the heterogeneous ES networks is necessary these days as many Systems developed using different networking protocols need to be interfaced and interconnected. The response time does not suffer. The electronics industry, as such, needs these kinds of solutions.

B. Rest of the Coverage in the Paper

In the rest of the paper, in Section 2, a review of the contributions made in the research related areas has been presented. In Section 3, application development using CAN and I^{2}C communication systems has been presented along with a comparison that shows how the networking systems differ and the issues related to interfacing between the communication systems. In Section 4, an architecture that focuses on the I^{2}C and CAN networks' hybridization has been presented. In Section 5, a computation method for determining optimum speeds of I^{2}C, CAN, and Ethernet has been presented, and conclusions have been drawn in Section 6.

II. Related Work

The speed of communication through I^{2}C and RS23 gets et influenced because of the complex electromagnetic conduit. Such a domain will result in the inappropriate disposal of the signals to the interfaces. Inappropriate transfer of electromagnetic will result in low speed and dependability. A few strategies are to be invented that help handle speed through legitimate cushion the executives [1].

Utilizing remote advances to improve conveyed implanted frameworks is testing contrasted with wired systems because of vulnerability and less unwavering quality caused by attractive obstruction, blurring, reflection, and so on. The consistency of remote correspondence is an issue. It turns out to be very unsafe for a framework when it needs to meet some basic security prerequisites by getting information through a remote system. The vulnerabilities existing with the remote correspondence can be settled through the utilization of mixed-race models or undertaking hybridization of structural models. The idea of hybridization is increasingly common in the automotive area where everything must work consummately, disregarding un- sureness. Hybridization helps when the correspondence needs to occur in questionable circumstances. Engineering has been displayed that can be utilized to actualize applications requiring the idea of hybridization. Numerous Interface gives that must be considered and taken care of through a programming model that makes the framework hybridization-mindful [2].

Two gauges are utilized as often as possible. They incorporate field bus and CAN bus for executing industry-based applications; field bus benchmarks are not uniform and incredibly vary from industry to industry. Correspondence between the gadgets utilizing Fieldbus in that capacity is entangled. In industry, both the transport based systems administration frameworks are often utilized. This has prompted a prerequisite of converting one kind of correspondence to the next, which can be accomplished through convention transformation. Convention transformation can be planned and actualized at the equipment level [3].

CAN transport based correspondence can be utilized for systems administration with frameworks. The engineers need to comprehend the CAN convention, Interface, controller, and Physical associations before the applications can be created utilizing CAN-based correspondence. Actualizing CAN-based correspondence at the net root level is entangled and needs thorough testing. The advancement of utilizations utilizing CAN is made to be straight through the CAN module. The CAN is a convention suite that can be coordinated with any inserted framework Software. Sending and getting the information can be accomplished through the CAN module [4]. When CAN is to be utilized alongside other correspondence conventions, for example, I2C, protocol transformations can be actualized at work level rather than net root level. If another module that modified works I2C communication is grown, at that point, convention transformation can be accomplished at the work level.

A different gadget can be structured and built to do conventional change utilizing numerous Microcontroller based frameworks, fast double port RAM information sharing innovation, and continuous multi- entrusting framework C/OS-II. A convention converter that helps correspondence somewhere in the range of RS232C and RS485 has been built up to be actualized inside savvy instruments, information securing frameworks, etc. [5]. The gadget can be interfaced with a remote checking framework through Ethernet-based correspondence. Along these lines, sequential gadgets can be connected to the system control layer. This gadget built up along these lines has unwavering high quality and continuous execution and acknowledge information trade, information sharing, and data handling among various Microcontroller based frameworks

Field bus convention is nonstandard and has been actualized in various forms. There is additionally an issue of interconnecting distinctive field buses. At the point when two systems are manufactured utilizing distinctive Fieldbus correspondence, conventional transformation is required. ARM controller can be used for accomplishing the change. The convention transformation is accomplished through the advancement of a protocol utilizing a standard information parcel. The strategy isn't constrained to coordinated transformation and is free of the transport area and convention utilized by the field buses [6].

A USB and I2C convention varies in numerous angles, considering how the correspondence is attempted. The information bundle groups, length of the system, number of gadgets that can be associated, number of ace, transport discretion, synchronization strategy, stream control, and so on vary greatly. A mapping between

I2C and USB have been done both at the equipment and product level. The product structure that can be utilized for accomplishing the change has been displayed [7].

Modbus and Profibus are two correspondence frameworks utilized for accomplishing modern mechanization. Both specialized strategies are generally utilized in the modern control field. Anyway, these two means of transport can't be associated straightforwardly because of the presence of extraordinary fluctuation between them. An entryway is required for interfacing two unique means of transport through which convention transformation can be conveyed. A passage is created utilizing the AT89C52 Microcontroller. Profibus and Modbus are two progressively basic mechanical field transport; they were generally utilized in the modern control field. Since the two means of transport can't inter-connect with one another, a Profibus and Modbus convention conversion is required. SPC3 is incorporated with the Micro Controller to achieve Profibus and Modbus conversions [8].

Numerous kinds of sub-frameworks are to be created and actualized, utilizing distinctive correspondence frameworks. For example, s Flight control, banking, therapeutic, and other high affirmation frameworks should be executed most unequivocally since the correspondence framework utilizes distinctive signaling, sheathing, commotion separating, signal disconnection, and so forth. One should structure and build up the framework so that one sub-framework doesn't meddle with the other.

Following a data stream at the equipment level is one strategy that can be utilized to distinguish and channel the differences. The door level in-arrangement stream following (GLIFT) framework is built to provide a technique for testing data streams inside I2C and USB. Time-division various access (TDMA) has been utilized that can confine a gadget on the BUS from the streams [9].

Mechanical Ethernet innovation (EPA) and Modbus correspondence innovation (MODBUS) are now and again utilized correspondence frameworks to actualize modern procedures. No immediate communication in that capacity can be conveyed in the middle of these two frameworks as there is no immediate similarity between them. A corresponding passage is created for accomplishing the necessary Interface between these two innovations. A passage has been created utilizing an ARM-based smaller scale controller and COS continuous working system. Bidirectional correspondence can be accomplished through the utilization of the entryway.

The correspondence entryway can give a steady, secure, constant, and adaptable answer to power plants [10].

Two kinds of industry explicit correspondence frameworks are utilized in the power segments intended for National power dispatching and control the breeze factories. To achieve integration, the framework theories correspondence frameworks must be interfaced with one another. A method has been developed to accomplish conventional transformation that executes capacities like interconnectivity, convention information type, configuration change, and scaling, information approval, the board of neighborhood/remote directions, recreation of information parcels transmission, etc. solicitations, communication bundles investigation, and repetitive correspondence joins [11].

Field buses are utilized for trading information between several microcontrollers and field gadgets by affecting communication among them. Numerous adaptations of the field bus communication frameworks exist. Structuring a correspondence framework for impacting correspondence among the gadgets that keep diverse Fieldbus correspondence benchmarks is mind-boggling and, by and large, prompts convoluted usage of equipment and programming. An effective correspondence interface is a requirement for executing a solid framework utilizing restrained field bus correspondence norms. CAN transport is additionally being utilized nowadays for actualizing a significant number of mechanical procedures. There is a need to interconnect between CANBUS and MODBUS. The correspondence conventions are to be mapped considering various parts of correspondence, and afterward, an interface is required to be created. An interface that associates both the transport based correspondence frameworks has been introduced [12][13].

The way networking of the embedded systems is carried on the kind of communication standard used I2C [14], USB [15], CAN [16] and RS485[17] and also the kind of Microcontroller based systems used for networking.

Many issues related to networking have been discussed [18][19][20][21][22][23][24] concerning testing Distributed Embedded Systems. Various methods and strategies have been proposed by Rajasekhar et al. [25] for hybridizing the networking of heterogeneous embedded networks. The interconnection between an I2C network and a CAN network can be achieved by developing a device that bridges both the networks. Speed Matching is one of the most important considerations that must be handled when it comes to the ES networks' hybridization [26]. Rajasekhar et al. have presented an efficient architecture that considers hybridization using Multi-master Interface [27], which is further extended in this paper to drive the architecture with the main considering driving through optimum speeds.

III. Application Development using Heterogeneous and Composite Embedded Network

A. Application Development using CAN Interface

An ES Application is developed through a CAN network, which is meant for monitoring and controlling temperature and humidity within an Engine.

CAN-based networking is archived through one Master and two slaves. The slaves are implemented through Arduino-UNO, and the Master is implemented through STEM 32. All three systems are connected through MCP 2515 CAN Module. The Master is connected to a SWITCH through its native Ethernet port for onward networking with another network developed using a different protocol such as I2C.

One of the slaves is connected with the DHT 11 sensor situated near an automobile system engine. The sensor continuously monitors the temperature of the engine and sends the information to the CAN master. CAN Master will send the information to the I2C Master through the Ethernet interface. I2C Master drives a different network based on its native protocol; The I^2C Master sends the data that it received from the CAN network to the I2C slave, a kind of actuator to cool through a FAN connected to the slave. The FAN is controlled through a relay system. Whenever the engine's temperature goes beyond a level, the FAN is switched on or otherwise switched off. The Functional requirements of CAN-based Application is shown in Table I.

The Networking Diagram for the Application is shown in Fig. 1.

CAN or Controller Area Network is a two-wired asynchronous, half-duplex fast sequential system-bus width of CAN is 217 bits. CAN is used for communication among devices in a closed distance, such as in a vehicle. CAN is based on CSMA-CD/ASM convention. CSMA guarantees that every hub must hang tight for a given period before sending any message. The crash location guarantees that the impact is kept away from choosing the messages dependent on their endorsed need. It gives a flagging rate from 125kbps to 1 Mbps. It accommodates 2048 diverse message identifiers.

TABLE I. CAN Application Description

Hardware Device Number	Hardware description	Interface description	Functional Description
1	STM32F401RE	CAN	To receive Temperature Data from Slave
			To receive Humidity Data from Slave
			To Send Temperature data to I^2C Master
			To Send Humidity data to I^2C Master
			To Receive Distance data from I2C Master
			To send distance data to the slave
2	ARDUINO UNO	CAN	To Receive distance data from Master
			To Control the Lighting System
3	ARDUINO UNO	CAN	To sense the temperature
			To sense the humidity
			To send Temperature to Master
			To send humidity to the Master

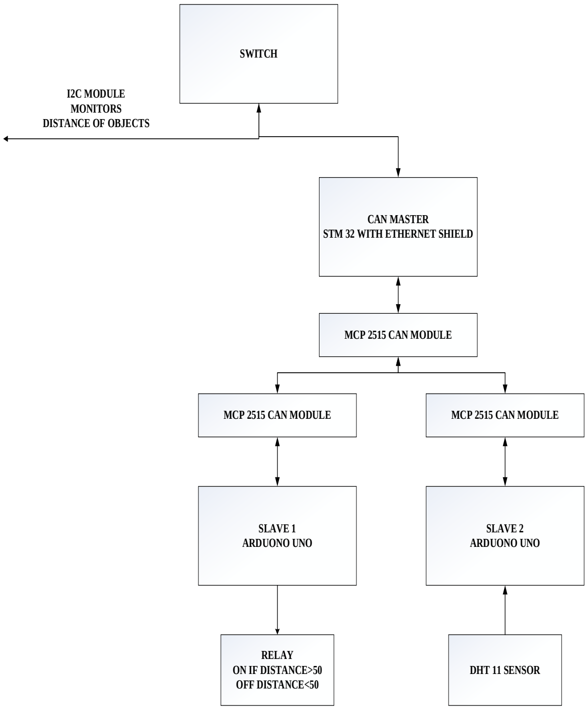


Fig. 1. CAN Networking Diagram.

CAN bus is the multi-master protocol. When the bus is idle, any device can be attached to the CAN bus and starts messaging. The can bus versatile, so devices attached to the bus do not have addressing. Each device in the CAN bus receives every message transmitted over the bus, and it is up to the device to decide whether to use the message that it receives or simply ignore it when the message is no more related to its own.

CAN bus provide remote transmission request (RTR), meaning that one node on the bus can request information from the other nodes. A request for information is sent to a node instead of waiting for a node to send information continuously. Any device in CAN bus can identify the error that occurred on the bus while transmitting the data and generates the error frames. The node which identifies the error alerts all other nodes about the error. There are no limitations for attaching and detaching the CAN bus devices, so devices are easy to attach and detach. Depending on the bus delays time and electrical loads, we can only decide the number of devices attached to the bus.

CAN protocol send messages in different types of data packets that include data frames, remote frames, error frames, and overload frames shown in Fig. 2, Fig. 3, Fig. 4, and Fig. 5. Data Frames are used to transmit data from Master to Slave and vice versa. Remote frames are used to seek permission from another node to transmit messages. Error frames are used for transmitting the errors that occur due to transmission, which can be classified as Bit errors, CRC errors, Form errors, Acknowledgement errors, and Stuffing errors. Overflow frames are used to create extra delays required for transmission of data or response.

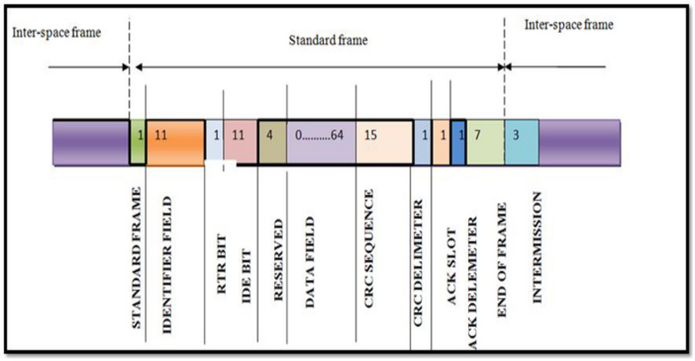


Fig. 2. Data Frame.

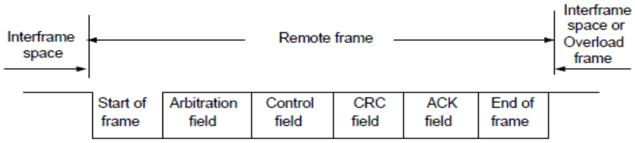


Fig. 3. Remote Frames.

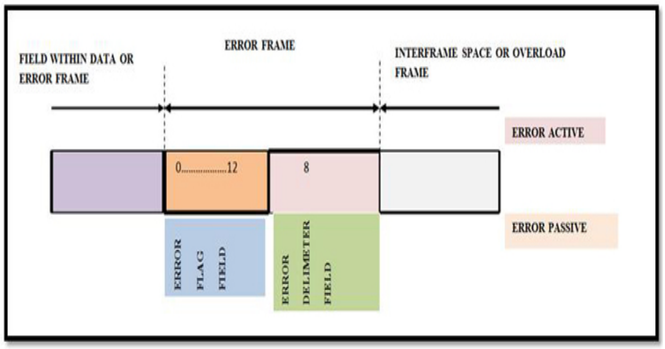


Fig. 4. Error Frame.

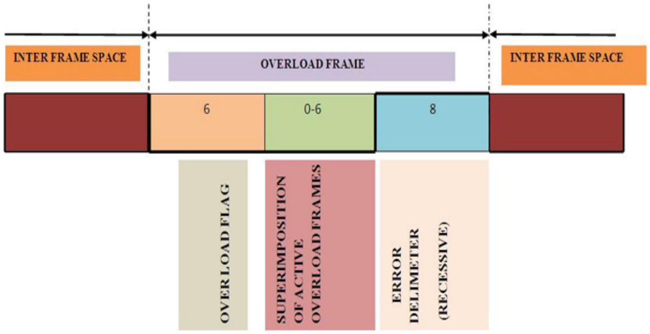

INTER FRAME SPACE
OVERLOAD FRAME
INTER FRAME SPACE
6
0-6
8
OVERLOAD FLAG
SUPERIMPOSITION OF ACTIVE OVERLOAD FRAMES
ERROR DELIMETER (RECESSIVE)

Fig. 5. Error Frame.

CAN-based communication between the Master and slave can be undertaken using speeds ranging from 125kbs to 1Mbps following the protocol sequences. The sequence in which the data packets are transmitted depends on the type of Application implemented over the network. The most appropriate choice of speed is the most important issue to be dealt with for realizing effective communication to take place among the heterogeneous network. The choice, however, needs to take into consideration many other important parameters.

B. *Application Development using I2C Interface*

Another I2C network is used as a subnetwork integrated into a composite network along with the CAN network. The ES Application developed through I2C Interface is related to measuring the distance of the objects Located behind a motor Vehicle and controlling the speed of a FAN fitted within the engine based on the engine's temperature, which is transmitted through the CAN network. The Application-specific functions implemented through I2C based networking is shown in Table II.

I2C Network is built with a single master and two slaves. One slave is connected with an ultrasonic sensor to monitor the nearest object's distance while the car is backing up. The monitored distances are sent to CAN master through Ethernet. The Second slave is connected with a DC Motor for controlling FAN's speed, which is connected to an engine. The FAN controlling is done based on the Temperature and Humidity data received from the CAN master. Two different protocols are used within I^{2}C, each for reading and writing data.

The networking Diagram for the I^{2}C based Application is shown in Fig. 6.

TABLE II. I2C APPLICATION DESCRIPTIONS

Hardware Device Number	**Hardware description**	**Type of Device**	**Functional Description**
1	**STM32F4 01 RE**	**Master**	To Receive Distance data from the slave
			To receive Temperature data from CAN Master
			To receive Humidity data from CAN Master
			To send Distance data to CAN Master
			To send Temperature data to the slave
			To send Humidity data to the slave
2	**STM 32 F301 RE**	**Slave**	To sense the distance of the object while reversing the car
			To send the Distance data to the Master
3	**STM 32 F301 RE**	**Slave**	To receive Temperature data from the Master
			To receive Humidity data from Master
			To actuate the DC motor for controlling the FAN

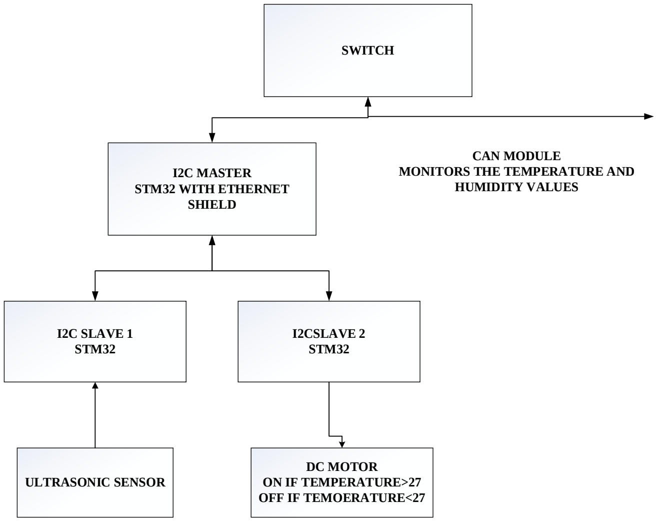

SWITCH
I2C MASTER STM32 WITH ETHERNET SHIELD
CAN MODULE MONITORS THE TEMPERATURE AND HUMIDITY VALUES
I2C SLAVE 1 STM32
I2CSLAVE 2 STM32
ULTRASONIC SENSOR
DC MOTOR ON IF TEMPERATURE>27 OFF IF TEMOERATURE<27

Fig. 6. I2C Networking Diagram.

The following sequence of operations is carried when data transmitted by a slave is to be read by the Master.

1) The master device sets the Read/Write bit to '1' instead of '0', which signals the targeted slave device that the master device is expecting data from it.

2) The slave device sends the 8 bits corresponding to the data block, and the master device sets the ACK/NACK bit.

3) Once the master device receives the required data, it sends a NACK bit. Then the slave device stops sending data and releases the SDA line.

4) Suppose the master device reads data from a specific internal location of a slave device. It first sends the location data to the slave device using the steps in the previous scenario. It then starts the process of reading data with a repeated start condition.

The following sequence of operations takes place when a master device tries to send data to a particular slave device through the I^{2}C bus:

1) The master device sends the start condition.

2) The master device sends the seven address bits, which corresponds to the slave device to be targeted.

3) The master device sets the Read/Write bit to '0', which signifies a write.

4) Now two scenarios are possible.

5) If no slave device matches with the address sent by the master device, the next ACK/NACK bit stays at '1' (default). This signals the master device that the slave device identification is unsuccessful. The master clock will end the current transaction by sending a Stop condition or a new Start condition.

6) If a slave device exists with the same address as the one specified by the master device, the slave device sets the ACK/NACK bit to '0', which signals the master device that a slave device is successfully targeted.

7) If a slave device is successfully targeted, the master device now sends 8 bits of data only considered and received by the targeted slave device. This data means nothing to the remaining slave devices.

8) If the slave device successfully receives the data, it sets the ACK/NACK bit to '0', which signals the master device to continue.

9) The previous two steps are repeated until all the data is transferred.

*10)*After all the data is sent to the slave device, the master device sends the Stop condition, which signals all the slave devices that the current transaction has ended.

C. Data Frames

I2C data is transferred in messages. Messages are broken into frames of data. Each message has an address frame that contains the binary address of the slave and one or more data frames that contain the data being transmitted. The message also includes start and stop conditions, read/write bits, and ACK/NACK bits between each data frame. The format of the message used within an I2C system is shown in Fig. 7.

Start Condition is initiated by making the SDA line switched from a high voltage level to a low voltage level before the SCL line switches from high to low. The Stop condition is achieved through switching the SDA line from LOW voltage to HIGH voltage after SCL is switched from LOW to HIGH. The Bits 7-10 contain the address of the slave with which the Master wants to communicate. The Read/Write Bit specifies whether the Master is sending data to the slave (low voltage level) or requesting it (high voltage level). The ACK/NACK Bit specifies whether the Master requires Acknowledgment from the slave or otherwise. If an address frame or data frame was successfully received, an ACK bit is returned to the sender from the receiving device.

D. Comparison of Application Specific CAN and I2C Networks

I2C and CAN networks differ in many ways: speeds, protocols used for transmission and reception of the data, addressing the devices within the networks, the data frames used for transmission and reception of the data, error control implemented, etc. Making a device as a slave to both the networks is cumbersome. There can be many ways of interconnecting both the networks, including connectivity through a single Master, Connectivity through Multiple Masters, Connectivity through a Bridge, and by implementing a Universal Bus.

E. Interconnecting between CAN and I2C Networks through Multi-master Interface

The interconnection between the I2C network and CAN network is achieved by interfacing the MASTERS using an Ethernet Interface is shown in Fig. 8.

The Master on the I2C network has both the I2C and Ethernet Interface, and similarly, the Master on the CAN network has both the CAN and Ethernet Interface. Whenever data from an I2C slave is transmitted to a CAN salve, the data is first transmitted to the I2C Master using the I2C protocol. The data packets are de-assembled and then assembled into Ethernet packets. The Ethernet packets are then transmitted to the CAN master through peer to peer connection established through Ethernet. The CAN-master receives the Ethernet packets, and the packets are dissembled and assembled into CAN packets, which are then transmitted to the CAN slave. The process of transmission from a CAN slave to the I2C slave similarly takes place.

The entire communication process involves selecting the proper speeds considering I2C, CAN, and Ethernet communication systems. The communication is completed with an acceptable response time. The delay caused due to de-assembling and assembling the packets must also be taken into account while calculating the response time.

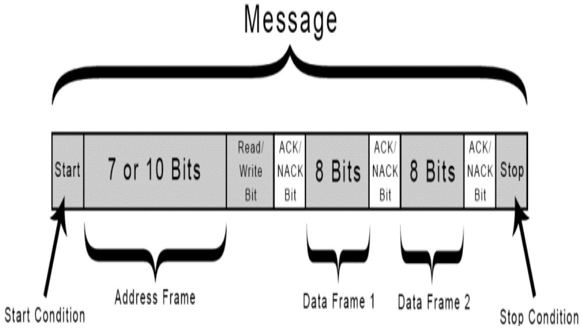


Fig. 7. I^{2}C Message Format.

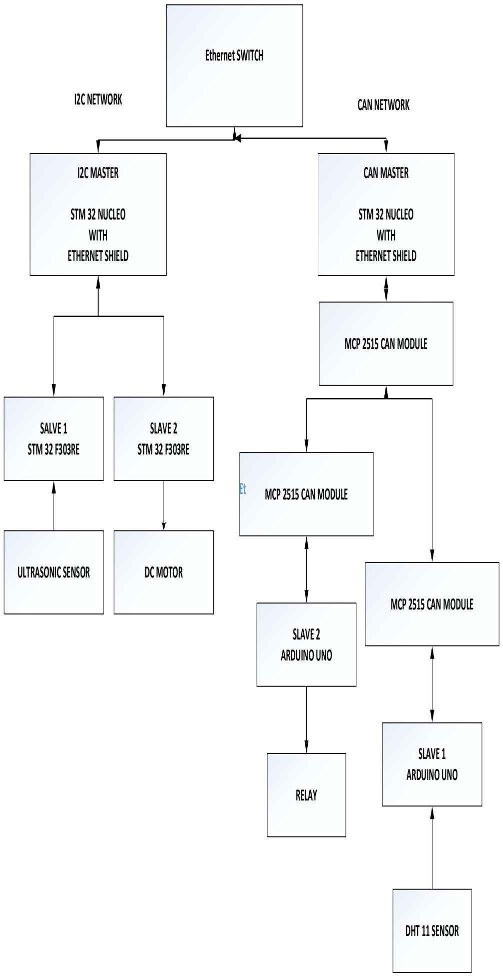


Fig. 8. Interconnecting I^{2}C and CAN Networks through Ethernet Interface.

IV. Architecture for Hybridised Communication in between I2C and CAN Networks

The architecture for establishing communication among I2C and CAN network through Ethernet-based Multi-master Interface is shown in Fig. 9. The most important issue is the arbitration among the I2C Master and CAN master on the kind of speeds used for effecting communications intra I2C, intra CAN, and between the masters using internet Interface. Speed matching is necessary so that the required response time is met. There should be time allowance for assembling and de-assembling the packets of different types. The software components contained within the slaves and the Masters shall carry designated functions as shown in the Architectural Diagram. The masters agree on speed based on the amount of data to be submitted, considering the type of packets and the packets' size to be transmitted, and considering the amount of delay time caused due to de-assembling and assembling processes. Once the speed agreements are achieved, the Master shall communicate the agreed speeds to the respective slaves to follow the speeds, especially setting the slave's internal timers.

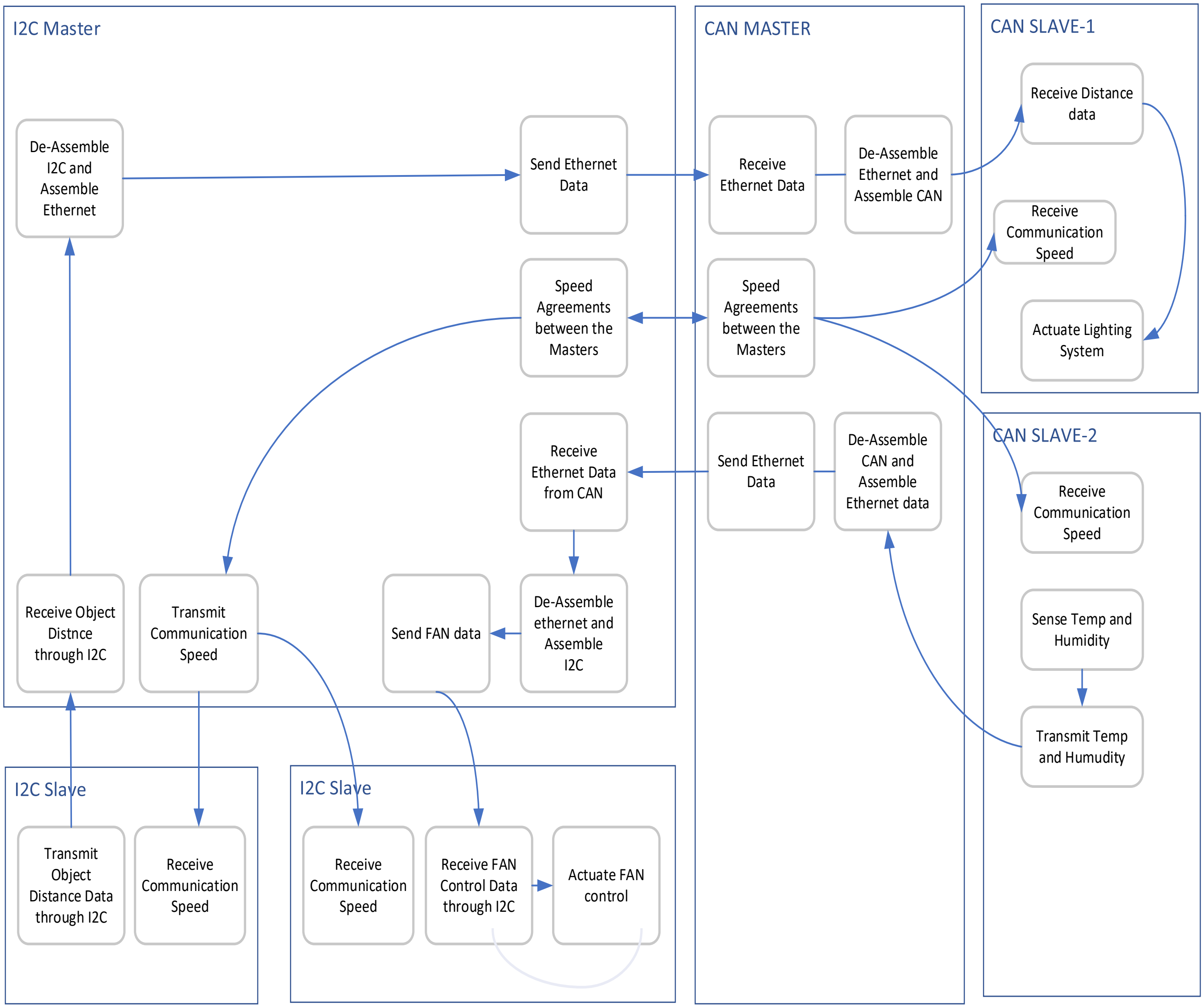

I2C Master
De-Assemble I2C and Assemble Ethernet
Send Ethernet Data
Speed Agreements between the Masters
Receive Ethernet Data from CAN
Receive Object Distnce through I2C
Transmit Communication Speed
Send FAN data
De-Assemble ethernet and Assemble I2C
CAN MASTER
Receive Ethernet Data
De-Assemble Ethernet and Assemble CAN
Speed Agreements between the Masters
Send Ethernet Data
De-Assemble CAN and Assemble Ethernet data
CAN SLAVE-1
Receive Distance data
Receive Communication Speed
Actuate Lighting System
CAN SLAVE-2
Receive Communication Speed
Sense Temp and Humidity
Transmit Temp and Humudity
I2C Slave
Transmit Object Distance Data through I2C
Receive Communication Speed
I2C Slave
Receive Communication Speed
Receive FAN Control Data through I2C
Actuate FAN control

Fig. 9. Architecture for Implementing MULTI-MASTER Ethernet-based Interface for Interconnecting I2C and CAN Networks.

V. Decision Making on Fixation of the Speeds of I2C, CAN and Ethernet for achieving Effective and Fast Communication

While data moves from one type of protocol to another, the data packet size increases or decreases. When an I2C Data packet is to be moved using the Ethernet protocol, the data packet size increases while the transmission speed increases; when a data packet is received through Ethernet into a CAN-based system, the same is de- pocketed and assembles into a packet size of small packets, which can be handled through fewer transmission speeds. It is rather challenging to decide on the transmission speeds chosen when communication has to be effected using a specific I2C, Ethernet, and CAN speeds that reduce the transmission time. The choice of speeds is also dependent on the total raw data transmitted from Time to Time.

In the typical example stated in section 5.0, the temperature data needs to be moved from the I2C network into the CAN network through the Ethernet interface. Similarly, the distance data must be moved in the other way. It is sufficient to analyze from either end of I2C and CAN for computing the time taken to transmit from either end. For this reason, the data analysis from I2C end to CAN is presented in this paper.

Communication time computations have been carried considering the data size that includes 16 Bits, 32Bits, 40Bits and 48Bits, I2C speeds that Include 100kbps, 400kbps, 3482kbps, and 5120kbps. CAN speeds include 500kbps, 250kbps, 125kbps, 10Kbps, and the Ethernet speeds that include 10240kbps, 20480kbps, 30720kbps, and 5120 Kbps to find the combination of speeds that provide the least response time. Table III, Table IV, Table V, and Table VI show the computations regarding data sizes 16bits, 32Bits, 40Bits, and 48Bits, respectively. The response time computations are made considering time taken to transmit using I2C protocol, data receiving time using Ethernet protocol, time taken to de-pocketing I2C packets and Pocketing to Ethernet packets, time is taken to transmit using Ethernet protocol, time is taken for receiving the data on the master side using the Ethernet protocol, time taken to DE packet the Ethernet packet to CAN packet, and time taken to transmit the CAN Packets to the CAN slave.

The response time computations are shown in Table VII, Table VIII, Table IX, and Table X for data sizes 16bits, 32Bits, 40Bits, and 48Bits for a different combination of speeds considering I2C, Ethernet, and CAN. It can be seen that the least response time (4.856002808 Micro Secs) is obtained when data size is 16 Bit when one considers I^{2}C speed of 5120 bits, Ethernet speed of 51200bits, and CAN speed of 500bits. The least response time obtained is 5.241206 Micro Seconds when data size is 32Bits and when one considers I2C speed of 3482bits, Ethernet speed of 51200bits, and CAN speed of 500bits. The least response time obtained is 4.517013550 microseconds when the data Size 40 Bits considering I2C speed 5120bits, Ethernet Speed of 20480bits, and CAN speed of 500bits. The least response time obtained is 5.542037964 Microseconds when the data size is fixed are 48bits considering the I2C speed being 5120bits, Ethernet speed being 51200bits, and CAN speed 500bits.

TABLE III. COMMUNICATION TIME COMPUTATIONS FOR DATA SIZE 16 BITS WITH DIFFERENT I2C, ETHERNET, AND CAN SPEEDS

I2C Packet Size in Bits	Number of I^{2}C Packets	Total Data to be Transmitted in Bits through I^{2}C	Speed in KBPS	Transmission Time Secs	Ethernet reconceiving Time in Secs	Ethernet De-Pocketing and Pocketing Time	Total Data to be Transmitted	Ethernet Speeds in KBPS	Time Taken to Transmit through Ethernet in Secs	Time Taken to Receive through Ethernet in Secs	Ethernet De-Pocketing and Pocketing Time	Total Data to be Transmitted in Bits	CAN Speed in KBPS	Time Taken to Transmit Data	Time Taken to Receive Data - CAN Slave Side	Time Taken to de pocketing the Packets on CAN Slave Size	Total Time Taken for Data Transmission
31	1	31	100	0.0003	0.00030	0.0001	512	10240	4.88E-05	4.88E-05	0.0001	87	500	0.00017	0.00017	0.0001	0.00080
			400	0.0001	0.00008	0.0001		20480	2.44E-05	2.44E-05	0.0001		250	0.00034	0.00034	0.0001	0.00073
			3482	0.0000	0.00001	0.0001		30720	1.63E-05	1.63E-05	0.0001		125	0.00068	0.00068	0.0001	0.00100
			5120	0.0000	0.00001	0.0001		51200	9.77E-06	9.77E-06	0.0001		10	0.0085	0.008496	0.0001	0.00870

TABLE IV. COMMUNICATION TIME COMPUTATIONS FOR DATA SIZE 32 BITS WITH DIFFERENT I2C, ETHERNET, AND CAN SPEEDS

I2C Packet Size in Bits	Number of I2C Packets	Total Data to be Transmitted in Bits through I2C	Speed in KBPS	Transmission Time Secs	Ethernet receiving Time in Secs	Ethernet De-Pocketing and Pocketing Time	Total Data to be Transmitted	Ethernet Speeds in KBPS	Time Taken to Transmit through Ethernet in Secs	Time Taken to Receive through Ethernet in Secs	Ethernet De-Pocketing and Pocketing Time	CAN Packet Suze in Bits	CAN Speed in KBMS	Time Taken to Transmit Data	Time Taken to Receive Data - CAN Slave Side	Time Taken to De pocketing the Packets on CAN Slave Size	Total Time Taken for Data Transmission
47	1	47	100	0.0005	0.00046	0.0001	512	10240	4.88281E-05	4.88281E-05	0.0001	103	500	0.000201172	0.000201172	0.0001	0.00098
			400	0.0001	0.00011	0.0001		20480	2.44141E-05	2.44141E-05	0.0001		250	0.000402344	0.000402344	0.0001	0.00083
			3482	0.0000	0.00001	0.0001		30720	1.6276E-05	1.6276E-05	0.0001		125	0.000804688	0.000804688	0.0001	0.00113
			5120	0.0000	0.00001	0.0001		51200	9.76563E-06	9.76563E-06	0.0001		10	0.010058594	0.010058594	0.0001	0.01027

TABLE V. COMMUNICATION TIME COMPUTATIONS FOR DATA SIZE 40 BITS WITH DIFFERENT I2C, ETHERNET, AND CAN SPEEDS

Total Data to be Transmitted in Bits through I2C	Speed in KBPS	Transmission Time Secs	Ethernet receiving Time in Secs	Ethernet De-Pocketing and Pocketing Time	Total Data to be Transmitted	Ethernet Speeds in KBPS	Time Taken to Transmit through Ethernet in Secs	Time Taken to Receive through Ethernet in Secs	Ethernet De-Pocketing and Pocketing Time	Total Data to be Transmitted in Bits	CAN Speed in KBMS	Time Taken to Transmit Data	Time Taken to Receive Data - CAN Slave Side	Time Taken to De pocketing the Packets on CAN Slave Size	Total Time Taken for Data Transmission
55	100	0.0005	0.00054	0.0001	512	10240	4.88281E-05	4.88281E-05	0.0001	111	500	0.000216797	0.000216797	0.0001	0.00108
	400	0.0001	0.00013	0.0001		20480	2.44141E-05	2.44141E-05	0.0001		250	0.000433594	0.000433594	0.0001	0.00088
	3482	0.0000	0.00002	0.0001		30720	1.6276E-05	1.6276E-05	0.0001		125	0.000867188	0.000867188	0.0001	0.00119
	5120	0.0000	0.00001	0.0001		51200	9.76563E-06	9.76563E-06	0.0001		10	0.010839844	0.010839844	0.0001	0.01105

TABLE VI. COMMUNICATION TIME COMPUTATIONS FOR DATA SIZE 48BITS WITH DIFFERENT I2C, ETHERNET, AND CAN SPEEDS

Total Data to be Transmitted in Bits through	Speed in KBPS	Transmission Time Secs	Ethernet receiving Time in Secs	Ethernet De-Pocketing and Pocketing Time	Total Data to be Transmitted	Ethernet Speeds in KBPS	Time Taken to Transmit through Ethernet in Secs	Time Taken to Receive through Ethernet in Secs	Ethernet De-Pocketing and Pocketing Time	Total Data to be Transmitted in Bits	CAN Speed in KBMS	Time Taken to Transmit Data	Time Taken to Receive Data - CAN Slave Side	Time Taken to De pocketing the Packets on CAN Slave Size	Total Time Taken for Data Transmission
63	100	0.000615234	0.000615234	0.0001	512	10240	4.88281E-05	4.88281E-05	0.0001	119	500	0.000232422	0.000232422	0.0001	0.00117
	400	0.000153809	0.000153809	0.0001		20480	2.44141E-05	2.44141E-05	0.0001		250	0.000464844	0.000464844	0.0001	0.00093
	3482	1.7671E-05	1.7671E-05	0.0001		30720	1.6276E-05	1.6276E-05	0.0001		125	0.000929688	0.000929688	0.0001	0.00126
	5120	1.20163E-05	1.20163E-05	0.0001		51200	9.76563E-06	9.76563E-06	0.0001		10	0.011621094	0.011621094	0.0001	0.01183

TABLE VII. RESPONSE TIME COMPUTATIONS WHEN DATA SIZE = 16 BITS

Response Time Computations when data Size = 16Bits		
I2C, Ethernet, CAN Speeds	**Response Time in Seconds**	**Normalized Response time**
100 ,1024 ,500	0.000821484375000	8.214843750
100 ,20480 ,500	0.000797070312500	7.970703125
100 ,30720,500	0.000788932291667	7.889322917
100, 51200,500	0.000782421875000	7.824218750
400,1024 ,500	0.000594433593750	5.944335938
400 ,20480 ,500	0.000570019531250	5.700195313
400 ,30720,500	0.000561881510417	5.618815104
400, 51200,500	0.000555371093750	5.553710938
3482,1024 ,500	0.000527445265826	5.274452658
3482 ,20480 ,500	0.000527445265826	5.274452658
3482 ,30720,500	0.000494893182493	4.948931825
3482, 51200,500	0.000488382765826	4.883827658
5120,1024 ,500	0.000485600280762	4.856002808
5120 ,20480 ,500	0.000500248718262	5.002487183
5120 ,30720,500	0.000492110697428	4.921106974
5120, 51200,500	0.000485600280762	4.856002808
100 ,1024 ,250	0.000991406250000	9.914062500
100 ,20480 ,250	0.000966992187500	9.669921875
100 ,30720,250	0.000958854166667	9.588541667
100, 51200,250	0.000952343750000	9.523437500
400,1024 ,250	0.000764355468750	7.643554688
400 ,20480 ,250	0.000739941406250	7.399414063
400 ,30720,250	0.000731803385417	7.318033854
400, 51200,250	0.000725292968750	7.252929688
3482,1024 ,250	0.000697367140826	0.697367141
3482 ,20480 ,250	0.000672953078326	0.672953078
3482 ,30720,250	0.000664815057493	0.664815057
3482, 51200,250	0.000658304640826	0.658304641
5120,1024 ,250	0.000694584655762	6.945846558
5120 ,20480 ,250	0.000670170593262	6.701705933
5120 ,30720,250	0.000662032572428	6.620325724
5120, 51200,250	0.000655522155762	6.555221558
100 ,1024 ,125	0.001331250000000	13.312500000
100 ,20480 ,125	0.001306835937500	13.068359375
100 ,30720,125	0.001298697916667	12.986979167
100, 51200,125	0.001292187500000	12.921875000
400,1024 ,125	0.001104199218750	11.041992188
400 ,20480 ,125	0.001079785156250	10.797851563
400 ,30720,125	0.001071647135417	10.716471354
400, 51200,125	0.001065136718750	10.651367188
3482,1024 ,125	0.001037210890826	10.372108908
3482 ,20480 ,125	0.001012796828326	10.127968283
3482 ,30720,125	0.001004658807493	10.046588075
3482, 51200,125	0.000998148390826	9.981483908
5120,1024 ,125	0.001034428405762	10.344284058
5120 ,20480 ,125	0.001010014343262	10.100143433
5120 ,30720,125	0.001001876322428	10.018763224
5120, 51200,125	0.000995365905762	9.953659058
100 ,1024 ,10	0.009147656250000	91.476562500
100 ,20480 ,10	0.009123242187500	91.232421875
100 ,30720,10	0.009115104166667	91.151041667
100, 51200,10	0.009108593750000	91.085937500
400,1024 ,10	0.008920605468750	89.206054688
400 ,20480 ,10	0.008896191406250	88.961914063
400 ,30720,10	0.008888053385417	88.880533854
400, 51200,10	0.008881542968750	88.815429688
3482,1024 ,10	0.008853617140826	88.536171408
3482 ,20480 ,10	0.008829203078326	88.292030783
3482 ,30720,10	0.008821065057493	88.210650575
3482, 51200,10	0.008814554640826	88.145546408
5120,1024 ,10	0.008850834655762	88.508346558
5120 ,20480 ,10	0.008826420593262	88.264205933
5120 ,30720,10	0.008818282572428	88.182825724
5120, 51200,10	0.008811772155762	88.117721558

TABLE VIII. RESPONSE TIME COMPUTATIONS WHEN DATA SIZE = 32 BITS

Response time Computations when Data size is 32Bits		
I2C, Ethernet, CAN Speeds	**Response time Seconds**	**Response time Normalized**
100 ,10240 ,500	0.00100898	10.089844
100 ,20480 ,500	0.00098457	9.845703
100 ,30720,500	0.00097643	9.764323
100, 51200,500	0.00096992	9.699219
400,10240 ,500	0.00066475	6.647461
400 ,20480 ,500	0.00084150	8.415039
400 ,30720,500	0.00063219	6.321940
400, 51200,500	0.00052412	5.241206
3482 ,10240 ,500	0.00056318	5.631831
3482 ,20480 ,500	0.00053877	5.387691
3482 ,30720,500	0.00053063	5.306311
3482, 51200,500	0.00052412	5.241206
5120,10240 ,500	0.00055896	5.589645
5120 ,20480 ,500	0.00053455	5.345505
5120 ,30720,500	0.00052641	5.264125

Response time Computations when Data size is 32Bits		
I2C, Ethernet, CAN Speeds	**Response time Seconds**	**Response time Normalized**
5120, 51200,500	0.00051990	5.199020
100 ,10240 ,250	0.00121016	12.101563
100 ,20480 ,250	0.00118574	11.857422
100 ,30720,250	0.00117760	11.776042
100, 51200,250	0.00117109	11.710938
400,10240 ,250	0.00086592	8.659180
400 ,20480 ,250	0.00084150	8.415039
400 ,30720,250	0.00083337	8.333659
400, 51200,250	0.00082686	8.268555
3482,10240 ,250	0.00076436	7.643550
3482 ,20480 ,250	0.00073994	7.399410
3482 ,30720,250	0.00073180	7.318029
3482, 51200,250	0.00072529	7.252925
5120,10240 ,250	0.00076014	7.601364
5120 ,20480 ,250	0.00073572	7.357224
5120 ,30720,250	0.00072758	7.275843
5120, 51200,250	0.00072107	7.210739
100 ,10240 ,125	0.00156133	15.613281
100 ,20480 ,125	0.00158809	15.880859
100 ,30720,125	0.00157995	15.799479
100, 51200,125	0.00157344	15.734375
400,10240 ,125	0.00126826	12.682617
400 ,20480 ,125	0.00124385	12.438477
400 ,30720,125	0.00123571	12.357096
400, 51200,125	0.00122920	12.291992
3482,10240 ,125	0.00116670	11.666988
3482 ,20480 ,125	0.00114228	11.422847
3482 ,30720,125	0.00113415	11.341467
3482, 51200,125	0.00112764	11.276363
5120,10240 ,125	0.00116248	11.624802
5120 ,20480 ,125	0.00113807	11.380661
5120 ,30720,125	0.00112993	11.299281
5120, 51200,125	0.01037732	103.773239
100 ,10240 ,10	0.01086641	108.664063
100 ,20480 ,10	0.01084199	108.419922
100 ,30720,10	0.01083385	108.338542
100, 51200,10	0.01082734	108.273438
400,10240 ,10	0.01052217	105.221680
400 ,20480 ,10	0.01049775	104.977539
400 ,30720,10	0.01048962	104.896159
400, 51200,10	0.01048311	104.831055
3482,10240,10	0.01042061	104.206050
3482 ,20480 ,10	0.01039619	103.961910
3482 ,30720,10	0.01038805	103.880529
3482, 51200,10	0.01038154	103.815425
5120,10240 ,10	0.01041639	104.163864
5120 ,20480 ,10	0.01039197	103.919724
5120 ,30720,10	0.01038383	103.838343
5120, 51200,10	0.01028709	102.870895

TABLE IX. RESPONSE TIME COMPUTATIONS WHEN DATA SIZE = 40 BITS

Response Time Computations when Data Size = 40Bits		
I2C, Ethernet, and CAN Speeds	**Response Time in Seconds**	**Normalized Response time**
100 ,10240 ,500	0.0011027344	11.027343750
100 ,20480 ,500	0.0010783203	10.783203125
100 ,30720,500	0.0010701823	10.701822917
100, 51200,500	0.0010636719	10.636718750
400,10240 ,500	0.0006999023	6.999023438
400 ,20480 ,500	0.0006754883	6.754882813
400 ,30720,500	0.0006673503	6.673502604
400, 51200,500	0.0006608398	6.608398438
3482,10240 ,500	0.0005810521	5.810520845
3482 ,20480 ,500	0.0005566380	5.566380220
3482 ,30720,500	0.0005485000	5.485000012
3482, 51200,500	0.0005419896	5.419895845
5120,10240 ,500	0.0005761154	5.761154175
5120 ,20480 ,500	0.0004517014	4.517013550
5120 ,30720,500	0.0005435633	5.435633341
5120, 51200,500	0.0005370529	5.370529175
100 ,10240 ,250	0.0013195313	13.195312500
100 ,20480 ,250	0.0012951172	12.951171875
100 ,30720,250	0.0012869792	12.869791667
100, 51200,250	0.0011804688	11.804687500
400,10240 ,250	0.0009166992	9.166992188
400 ,20480 ,250	0.0008922852	8.922851563
400 ,30720,250	0.0008841471	8.841471354
400, 51200,250	0.0008776367	8.776367188
3,48,21,02,40,250	0.0007978490	7.978489595
3482 ,20480 ,250	0.0007734349	7.734348970
3482 ,30720,250	0.0007652969	7.652968762
3482, 51200,250	0.0007587865	7.587864595
5120,10240 ,250	0.0007929123	7.929122925
5120 ,20480 ,250	0.0007684982	7.684982300
5120 ,30720,250	0.0007603602	7.603602091

Response Time Computations when Data Size = 40Bits		
I2C, Ethernet, and CAN Speeds	**Response Time in Seconds**	**Normalized Response time**
5120, 51200,250	0.0007538498	7.538497925
100 ,10240 ,125	0.0017531250	17.531250000
100 ,20480 ,125	0.0017287109	17.287109375
100 ,30720,125	0.0017205729	17.205729167
100, 51200,125	0.0017140625	17.140625000
40,01,02,40,125	0.0013502930	13.502929688
400 ,20480 ,125	0.0013258789	13.258789063
400 ,30720,125	0.0013177409	13.177408854
400, 51200,125	0.0013112305	13.112304688
3482,10240 ,125	0.0012314427	12.314427095
3482 ,20480 ,125	0.0012070286	12.070286470
3482 ,30720,125	0.0011988906	11.988906262
3482, 51200,125	0.0011923802	11.923802095
5120,10240 ,125	0.0012265060	12.265060425
5120 ,20480 ,125	0.0012020920	12.020919800
5120 ,30720,125	0.0011939540	11.939539591
5120, 51200,125	0.0011874435	11.874435425
100 ,10240 ,10	0.0117257813	117.257812500
100 ,20480 ,10	0.0117013672	117.013671875
100 ,30720,10	0.0116932292	116.932291667
100, 51200,10	0.0116867188	116.867187500
400,10240,10	0.0113229492	113.229492188
400 ,20480 ,10	0.0112985352	112.985351563
400 ,30720,10	0.0112903971	112.903971354
400, 51200,10	0.0112838867	112.838867188
3482,10240 ,10	0.0112040990	112.040989595
3482 ,20480 ,10	0.0111796849	111.796848970
3482 ,30720,10	0.0111715469	111.715468762
3482, 51200,10	0.0111650365	111.650364595
5120,10240 ,10	0.0111991623	111.991622925
5120 ,20480 ,10	0.0111747482	111.747482300
5120 ,30720,10	0.0111666102	111.666102091
5120, 51200,10	0.0111503342	111.503341675

TABLE X. RESPONSE TIME COMPUTATIONS WHEN DATA SIZE = 48 BIT

Response time Computation when the Data Size = 48Bits		
I2C, Ethernet, CAN Speed	**Response time in Seconds**	**Normalized Response time**
100 ,10240 ,500	0.0011964844	11.964843750
100 ,20480 ,500	0.0011720703	11.720703125
100 ,30720,500	0.0011639323	11.639322917
100, 51200,500	0.0011574219	11.574218750
400,10240 ,500	0.0007350586	7.350585938
400 ,20480 ,500	0.0007106445	7.106445313
400 ,30720,500	0.0007025065	7.025065104
400, 51200,500	0.0006959961	6.959960938
3482,10240 ,500	0.0005989210	5.989210241
3482 ,20480 ,500	0.0005745070	5.745069616
3482 ,30720,500	0.0005663689	5.663689408
3482, 51200,500	0.0005598585	5.598585241
5120,10240 ,500	0.0005932663	5.932662964
5120 ,20480 ,500	0.0005688522	5.688522339
5120 ,30720,500	0.0005607142	5.607142131
5120, 51200,500	0.0005542038	5.542037964
100 ,10240 ,250	0.0014289063	14.289062500
100 ,20480 ,250	0.0014044922	14.044921875
100 ,30720,250	0.0013963542	13.963541667
100, 51200,250	0.0013898438	13.898437500
400,10240 ,250	0.0009674805	9.674804688
400 ,20480 ,250	0.0009430664	9.430664063
400 ,30720,250	0.0009349284	9.349283854
400, 51200,250	0.0009284180	9.284179688
3482, 10240, 250	0.0008313429	8.313428991
3482 ,20480 ,250	0.0008069288	8.069288366
82 ,30720,250	0.0007987908	7.987908158
3482, 51200,250	0.0007922804	7.922803991
5120,10240 ,250	0.0008256882	8.256881714
5120 ,20480 ,250	0.0008012741	8.012741089
5120 ,30720,250	0.0007931361	7.931360881
5120, 51200,250	0.0007866257	7.866256714
100 ,10240 ,125	0.0018937500	18.937500000
100 ,20480 ,125	0.0018693359	18.693359375
100 ,30720,125	0.0018611979	18.611979167
100, 51200,125	0.0018546875	18.546875000
400, 10240, 125	0.0014323242	14.323242188
400 ,20480 ,125	0.0014079102	14.079101563
400 ,30720,125	0.0013997721	13.997721354
400, 51200,125	0.0013932617	13.932617188
3482,10240 ,125	0.0012961866	12.961866491
3482 ,20480 ,125	0.0012717726	12.717725866
3482 ,30720,125	0.0012636346	12.636345658
3482, 51200,125	0.0012571241	12.571241491
5120,10240 ,125	0.0012905319	12.905319214
5120 ,20480 ,125	0.0012661179	12.661178589
5120 ,30720,125	0.0012579798	12.579798381

Response time Computation when the Data Size = 48Bits		
I2C, Ethernet, CAN Speed	**Response time in Seconds**	**Normalized Response time**
5120, 51200,125	0.0012514694	12.514694214
100 ,10240 ,10	0.0125851563	125.851562500
100 ,20480 ,10	0.0125607422	125.607421875
100 ,30720,10	0.0125526042	125.526041667
100, 51200,10	0.0125460938	125.460937500
400,10240,10	0.0121237305	121.237304688
400 ,20480 ,10	0.0120993164	120.993164063
400 ,30720,10	0.0120911784	120.911783854
400, 51200,10	0.0120846680	120.846679688
3482,10240 ,10	0.0119875929	119.875928991
3482 ,20480 ,10	0.0119631788	119.631788366
3482 ,30720,10	0.0119550408	119.550408158
3482, 51200,10	0.0119485304	119.485303991
5120,10240 ,10	0.0119819382	119.819381714
5120 ,20480 ,10	0.0119575241	119.575241089
5120 ,30720,10	0.0119493861	119.493860881
5120, 51200,10	0.0119428757	119.428756714

TABLE XI. COMPARATIVE ANALYSIS OF I2C, ETHERNET AND CAN SPEEDS @ DIFFERENT DATA SIZES

Data Size	**I2C, Ethernet, CAN speeds**	**Response time in Seconds**	**Response Time in Micro Seconds**
16	5120, 51200,500	0.00048560	4.85600
32	3482, 51200,500	0.00052412	5.24121
40	5120 ,20480 ,500	0.00045170	4.51701
48	5120, 51200,500	0.00055420	5.54204

The overall assessments of comparable speeds considering local minimums of different data sizes and their related speeds are shown in Table XI. It can be seen from Table XI and Fig. 10 that optimum response time can be achieved when I2C speed is 5120 Bits, Ethernet speed is fixed at 20480 Bits, and the CAN speed fixed at 40bits and that too when the data size is fixed at 40 Bytes. However, if there is a limitation on the data size, then the local minimum can be determined considering all the combinations of the I2C, Ethernet, and CAN speeds.

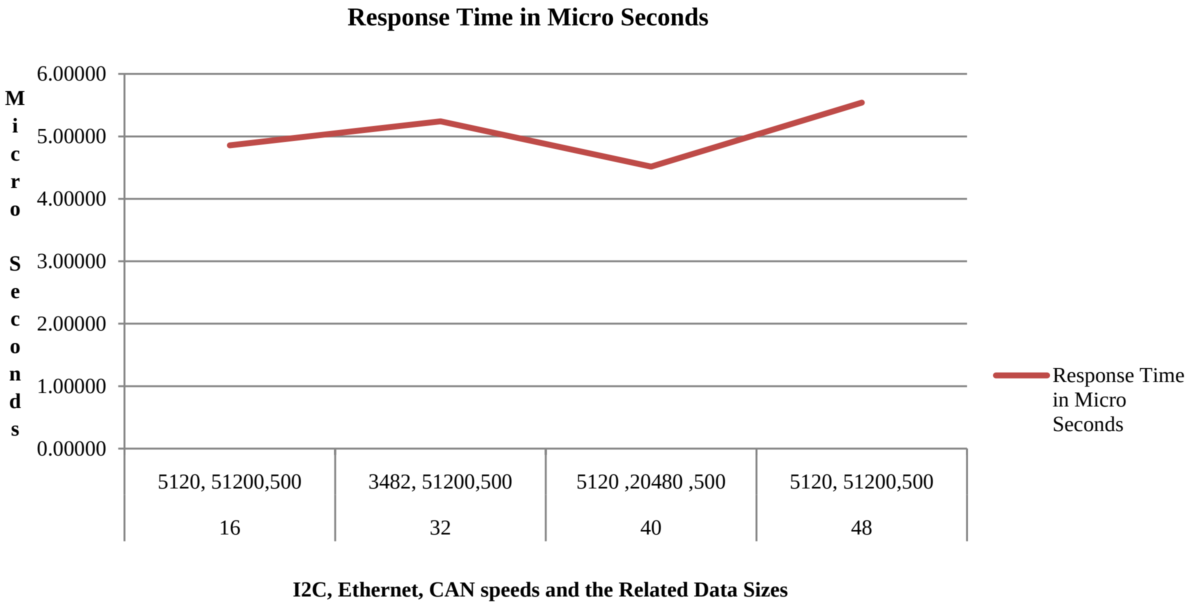


Fig. 10. Optimum Response Time among Many Optimum Locals.

VI. Conclusions

Interfacing different heterogeneous ES networks into a composite network are quite challenging. Many issues that include synchronizing, arbitration, timing, and speed control have to be tackled properly so that optimum acceptable response time is achieved, failing which the system's expected response time cannot be supported. Many mechanisms can be implemented for achieving the issues of hybridization that include introducing a bridge device, adapting a universal bus, a single master interfacing, and a Multi-Master interfacing. A detailed comparison of protocols shows the issues that must be considered for interacting and integrating heterogeneous communication systems.

In this paper, a Multi-Master interfacing interconnects an I2C network with a CAN network using an Ethernet interface. The detailed working of data communication considering the different combination of I2C, CAN, and Ethernet speeds have been presented, and the best combination of these speeds that gives the least response time has been presented.

Further research can be carried to find the speed combinations that must be considered when networking is to be done considering I2C+USB, I2C+RS485, CAN+USB, CAN+RS485, USB_RS485. Interfacing using a single master through a common bus is also one of the important approaches that can be considered.

References

[1] CHAI Yan-Jie, SUN Ji-yin, GAO Jing, TAO Ling-jiao, JI Jing, BAO Fei-hu," Improvement of I2C Bus and RS-232 Serial Port un-der Complex Electromagnetic Environment" International Conference on Computer Science and Software Engineering, 2008, PP 178 - 181.

[2] Antonio Casimiro, Jose Rufino, Luis Marques, Mario Calha, and Paulo Verissimo "Applying Architectural Hybridization in Net-worked Embedded Systems," IFIP International Federation for In-development Processing, 2009, PP 264-275.

[3] Lou Guohuan, ZhangHao, Zhao Wei, "Research on Designing Method ofCAN Bus and Modbus Protocol Conversion Interface" International Conference on Future BioMedical Information Engineering,2009, PP 180-182.

[4] Xiaoming Li, Mingxiong Li, "An Embedded CAN-BUS Communication Module for Measurement and Control System" International conference on ICEEE, 10.119/ICEEE.2010.5661248, 2010.

[5] Peng Daogang, Zhang Hao, Li Hui, Xia Fei," Development of the Communication Protocol Conversion Equipment Based on Embedded Multi-MCU and μC/OS-II," International Conference on Measuring Technology and Mechatronics Automation, 2010, PP 15-18.

[6] Tae-Won Kim, Jin Ho Kim, Do Eon Lee, Jun young Moon, Jae Wook Jeon," Development of Gateway based on BroadR-Reach for Application in Automation Network," International Conference on Computing, Communication and Automation, 2016, PP 421-426.

[7] CHEN Huijuan," Heterogeneous Network Integration Based on Protocol Conversion" Control Conference (CCC), 2016, PP 6888-6893.

[8] Chen Wei, Wang Xijun, Sun Wenxia," The Design of profinet- Modbus protocol conversion Gateway Based on the ERTEC 200P", International Conference on Software, Knowledge, Information Management & Applications, 2016, PP 87-91.

[9] Jason Oberg, Wei Hu, Ali Irturk," Information Flow Isolation in I2C and USB", Design Automation Conference, 2011, PP 254-259.

[10] Li Hui, Zhang Hao, Peng Daogang," Design and Application of Communication Gateway of EPA and MODBUS on Electric Power System" International Conference on Future Electrical Power and Energy Systems, 2012, PP 286 – 292.

[11] Gheorghe G.Florea, Oana, Rohat, Monica G. Dragan, "Contributions to the Development of Communication Protocols Conversion Equipment," 2nd IFAC Workshop on Convergence of Information Technologies and Control Methods with Power Systems, 2013, VOLUME 46, Issue 6, PP 84-88.

[12] Umesh Goyal, Gaurav Khurana," Implementing MOD bus and CAN bus Protocol Conversion Interface" IJETT, 2013, VOLUME 4, Issue 4, PP 630-635.

[13] Roopak Sinha," Conversing at Many Layers: Multi-layer System-On-chip Protocol Conversion," 20th International Conference on Engineering of Complex Computer Systems, 2015, PP 170-173.

[14] JKRSastry, J. Viswanadh Ganesh, and J. SasiBhanu, "I2C based Networking for Implementing Heterogeneous Microcontroller based Distributed Embedded Systems", Indian Journal of Science and Technology, 2015, Vol 8(15), PP 1-10.

[15] JKRSastry, Valluru Sai Kumar Reddy, Smt J SasiBhanu," Net-working Heterogeneous Microcontroller based Systems through Universal Serial Bus," IJECE, 2015, Vol. 5, No. 5, PP. 992-1002.

[16] JKR. Sastry, M. Vijaya Lakshmi, and Smt J. SasiBhanu," Optimizing Communication Between Heterogeneous Distributed Embedded Systems Using CAN Protocol," ARPN Journal of Engineering and Applied Sciences, 2015, VOL. 10, NO. 18, PP 7900-7911.

[17] JKR Sastry, T. Naga Sai Tejasvi and J. Aparna, Dynamic scheduling of message flow within a distributed embedded system connected through RS485 network, ARPN Journal of Engineering and Applied Sciences, VOL. 12, NO. 9, MAY 2017.

[18] J. K. R. Sastry, A. Suresh, and Smt J. Sasi Bhanu, Building Heterogeneous Distributed Embedded Systems through RS485 Communication Protocol, ARPN Journal of Engineering and Applied Sciences, issue. 16, vol.10, 2015.

[19] K. Chaitanya, Sastry JKR, K. N. Sravani, D. Pavani, Ramya, and K. Rajasekhara Rao, Testing Distributed Embedded Systems Using Assert Macros, ARPN Journal of Engineering and Applied Sciences, 2017, page no.3011-3021.

[20] Sastry JKR, K. Chaitanya, K. Rajasekhara Rao, DBK Kamesh, Testing Distributed Embedded Systems Through Instruction Set Simulators, PONTE, International Journal of Sciences and Research, issue.7, vol.73, July 2017, page no.353-382.

[21] JKR Sastry, K. Chaitanya, K. Rajasekhara Rao, DBK Kamesh, An Efficient Method for Testing Distributed Embedded Systems using In-circuit Emulators, PONTE, International Journal of Sciences and Research, issue.7, vol.73, 2017, page no.390-422.

[22] K. Chaitanya, JKR Sastry, K. Rajasekhara Rao, Testing Distributed Embedded Systems Using Logic, Analyzer, International Journal of Engineering and Technology, March 2018, page no. 297-302.

[23] K Chaitanya1, Dr. K Rajasekhra Rao, Dr. JKR Sastry, A Framework for Testing Distributed Embedded Systems, International Journal advanced Trends in computer science and engineering, 2019, Volume 8, No.4, PP. 1104-1227.

[24] K Chaitanya, Dr. K Rajasekhra Rao, Dr. JKR Sastry. A Formal and Enriched Framework for Testing Distributed Embedded Systems, International Journal Emerging Trends in Engineering Research, Volume 7, No. 12, 2019, PP. 867-878.

[25] J. Rajasekhar, Dr. JKR Sastry, An approach to hybridization of embedded system networks, International Journal of Engineering & Technology, 7 (2.7) (2018) 384-389.

[26] Jammalamadaka Rajasekhar, JKR Sastry, Building Composite Embedded Systems Based Networks Through Hybridization and Bridging I2C and CAN, Journal of Engineering Science and Technology, Vol. 15, No. 2 (2020) 858 – 88.

[27] E. Mounika, Dr. JKR Sastry, J. Rajasekhar, A Hybridised Heterogeneous Embedded System Networking through Multi-Master Interface, International Journal of Emerging Trends in Engineering Research, Volume 8. No. 3 March 2020, https://doi.org/10.30534/ijeter/2020/45832020.

An Analysis of Human Activities Recognition using Smartwatches Dataset

Saadia Karim[1], SM Aqil Burney[2], Nadeem Mahmood[3]
College of Computer Science and Information System, IoBM, Karachi, Pakistan[1, 2]
Umaer Basha Institute of Information Technology, University of Karachi, Karachi, Pakistan[3]

***Abstract*—Today, the era of smart devices evolving the human behavior interaction to a changing environment where the learning of activities is monitored to predict the next step of human behavior. The smart devices have these sensors built-in (accelerometer and gyroscope), which are continuously generating a large amount of data. The data used to identify the novel patterns of human behavior, together with machine learning and data mining techniques. Classification of human motions with motion sensor data is among the current topics of study. The classification is an important part of data mining techniques and used in this work to find the accuracy of instances in the given dataset. Thus, it is possible to follow the activities of a user carrying only a smartwatch. The smartwatches consisting of four different models from two manufacturers are used. Furthermore, the experiment contains nine users and seven activities performed by them. After the classification was determined, the data set to which the principal component analysis has been applied was classified by decision stump, j48, Bayes net, naive Bayes, naive Bayes multinomial text, random forest, and logit boost methods, and their performances were compared. The most successful result was obtained from the random forest method. The accuracy of the Random Forest classification algorithm on nominal datasets is 99.99% on both accelerometer and gyroscope sensors.**

***Keywords*—Human activity recognition; smartwatches; big data; machine learning; random forest**

I. Introduction

In this technology era, various advanced applications are developing to support human daily activities. Among them, smartphones, and smartwatches are more popular to use. Smartwatches are wearable computer in the form of a wristwatch. Smartwatches are providing a touchscreen interface for daily use and become an integrated part of everyday life [1] [2]. They frequently use smartwatches producing huge amounts of data on daily basis. The growing era of technology, make smartwatches more power full device because it comprises the efficient computational power, internet facilities, and hardware sensors such as accelerometer [3][4][5], gyroscope [3][6], Face lock mechanism, touch with finger lock, and GPS receivers in it [7]. All these topographies inspire to study human activity recognition system and make the smartwatches to become a rich environment for many systems like:

- Blood pressure and temperature measuring in COVID19 situation [8].
- Heterogeneities Activity Recognition [9].
- Activity Recognition Using Active Learning [1].
- Classification of Various Daily Activities [10].
- Sport-Related Human Activity Detection and Recognition [11].

Smartwatches sensors become monitoring tools to recognize human physical activities such as Biking, Walking, Stair Up, Stairs down, Sitting, Standing, and doing nothing (null). Analyzing and learning of these activities will help to monitor human health and provide health services, and security services.

Serkan with other authors [3] describes the motion detection of human daily life, is an important task such as for healthcare, fitness, children's motion, and older person care. The smartwatches are easy to carry on hand. So, the activities of human daily life are observed in the form of data set through smartwatches. The eight different activities are collected for analysis to be performed. The analysis used the machine learning approach. The features are determined by using random forest, support vector machine, C4.5, and k-nearest neighbor methods. The most successful result was obtained from the random forest.

Hamid and Ali in [7] used smartwatches, which were fixed on the right human ankle-foot. The accelerometer and gyroscope sensor are used to classify the walk, jog, and run base on fast, normal, and slow motion. A threshold-based analysis using 20 different human activities are classified. The accuracy of 97.5% is achieved from the raw dataset of smartwatches used in the experiment. The authors in [9] used two different brands of smartwatches, which were fixed on the hand. The smartwatches contain different types of embedded sensors for example accelerometer, Global positioning system (GPS), gyroscope, Wi-Fi, and NFC. The output from these sensors has been used for the calculation of accurate sensing and impact on HAR by Sensor Biases (SB), Sensor Rate Heterogeneity (SRH), and Sampling Rate Instability (SRI). The clustering approach (low-pass filtering) has been used for finding the justification deficiencies in HAR systems. The 36 devices have been used for getting datasets and examination identifies the type heterogeneity has present in the datasets. The experiments show that the sensor biases have an 8% deviation in it, whereas the heterogeneity in data causes a huge challenge for monitoring the tasks. There is the time domain, feature domain and ECDF types of feature extraction has used with four types of learning approaches: C4.5, SVM, K-NN, and random forests. The author adopts the F1-score, which is the harmonic mean of precision and recall, as the primary evaluation metric.

The focus of this article is to classify the given dataset on the performance evaluation of selected classifiers. The method used in this article are quite straightforward, but to generate results that are competitive with available articles. This article is organized as follows. Section 2 describes the related work already done for smartwatches. Section 3 then describes the process for collecting the smartwatches sensor data and transforming it into a form suitable for data mining. Section 4 then describes the performance evaluation based on accuracy and time using classifiers on smartwatches datasets with the results from applying those classifiers. The article concludes with Section 5, which summarizes the main conclusions, identifies areas for future work.

II. Related Work

Akram and other authors in [12] illustrate that recognizing human physical activities are a significantly challenging research region for the researcher to work and many new applications are functioning as healthcare, smart cities, and home security. The new capable method is to track human activities by adding sensors to smartphones. The model is proficient in identifying multiple activities of humans in a real-time environment achieved dataset daily by triaxial accelerometer construct in smartphones. The experiment does not work on some constraints like fast walking, dancing. The design of the system works on the new digital low-pass filter with some classification methodology on high and low speed of data to make the system more accurate and faster. The selected classifiers are available in the Weka toolkit: Multilayer Perception, Random Forest, SVM, Simple Logistic, Logit Boost, and LMT. To improve the accuracy, proficiency, and toughness of classifiers, the fusion method has been used to combine different classifiers and found the average probability accurate rate of 91.15% [12].

Whereas, in [13], the authors express the Weka tool for data mining purposes. Data mining involves the following steps as data preprocessing, cleaning, transformation, reduction, feature selection, classification, and evaluation of data. The Weka tool is used to process all steps involves in data mining. The Weka tool accepts the ARFF format and CSV format. The dataset loads in Weka to be preprocessed and apply the classifications methods, which help predictions and apply algorithm techniques. Classification, clustering, and association rules are discussed to get interaction and give an overview of using the Weka tool [13].

According to [14], the authors describe the advanced method of monitoring human activities by an accelerometer, Global positioning system (GPS), gyroscope, radio frequency identification, and NFC available in smartphones using discrete data variables. The enabling of automatic sensors extend the time for ubiquitous computing. The data collected from devices are analyzed by some classification methods then perform monitoring by some algorithms like Bayesian decision (BDM), support vector machines (SVM), K-nearest neighborhood (KNN), etc. The usage of different devices and these devices has different consumption of batteries. This energy consumption is a major problem by smartphone IMU sensors to execute all tasks needed for activity classifications. The dynamic Ameva algorithm has been applied to monitor 32Hz sitting, standing, and lying to 50Hz of walking. Ameva classification system has a 98% average accuracy for recognizing different activities. To identify the accuracy in heterogeneous activities the Dynamic Sample Rate and Duty Cycle section approaches are used. The system analysis of smart activities by items is additional information to provide support to the system. The comparison of work has been performed with other classification approaches and concluded that all approaches give related results but the computational cost is applied [14]. As smartwatches are used to recognize the activities of humans. In [1], the author elaborates on the usage of smartwatches based on hardware and software sensors and active learning mechanisms. The hardware and software-based sensors using accelerometer, rotation vector, and linear acceleration sensors. These sensors produce data for 5 daily activities perform by users such are standing, sitting, laying down, walking, and none. These activities are classified by Random Forest, Extra Tree, Naive Bayes, Logistic Regression, and Support Vector Machine (SVM). Among them, the Extra Tree model is used as a baseline for the active learning model. Through the active learning model, 95% of accuracy is proved using 46% fewer samples [1]. Furthermore, in [10], the author proof 97.19% accuracy using a deep learning model through a Convolutional Neural Network (CNN). The classification uses 11 different activities (high or less use in daily life). This study shows that if accuracy measures are predicted accurately then the smart-home concept saves energy and this will help us to go the concept of IoT.

In addition to the field of smartwatches and human activity recognition, Weiss with other authors [15] added biometric efficiency to the devices. For evaluating the 18 different activities in the WISDM lab dataset. The biometric authentication in smartwatches and smartphones gives zero-effort feasible results and help to identify the subject and apply access to them. The IoT and energy saving using smartwatches same as smartwatches used for monitoring player health. The activities performed by the player have many variations and recognize them is a complex task.

The author in [11], explains the complication of monitoring player activities on the sliding window method as candidate motion and action changes rapidly. The activities related to sports are monitored based on the duration of each task to be done within an interval of time and these intervals of time are classified by the CNN technique. The CNN technique is based on the periodic and non-periodic duration of activities. This helps to identify the weak and strong candidates in the team as compared to the previous sliding window method. Same as in [8], Rozita Jamili with other authors explains the importance of IoT during the pandemic era of Corona Virus Disease (COVID-19). There are many sensors used in smart-home among them smartwatches are a reliable source of monitoring human activities. The smartwatches on the human hand enable us to monitor the health and take the required action as soon as possible. The method includes the MQTT, Web Socket, and HTTP programming to access and secure the devices. The data collected for the emergency of the patient gives 95% accuracy, which helps to provide good health services from the health department [8].

III. Methodology

This section focused on step by step procedures perform on different activities using accelerometer and gyroscope data collected from smartwatches. First, we need to collect data then. Secondly, we describe the tool with which the analysis will perform. Then which type of classifiers are used to extract information from the data. Finally, we use the analysis to justify the accuracy of the classifier used in the experiment and draw the comparative analysis.

A. Data Collection

For data collection. The dataset is downloaded from the UCI machine learning repository [9] [16]. UCI machine learning repository is a gathering of databases that are used by the machine learning community for the experimental examination of machine learning procedures. The data from the accelerometer and gyroscope includes 10 attributes among them the motion of smartwatches is shown by x, y, and z axes. The dataset has nine different users as shown in Table I. The four smartwatches from two different brands (Samsung and LG as shown in Fig. 1 are used for recording the seven different activities of humans.

B. Weka Tool

Weka is a collection of machine learning algorithms and preprocessing virtual tools for different purposes of data mining, written in Java, and developed at the University of Waikato, New Zealand. Weka has multiple learning algorithms with various transformations on datasets. The datasets can be preprocessed, captured into a learning schema to analyze result-using classifiers without writing any code. Weka provides filters for preprocessing, classifiers, clusters, association, and visualization in 2D and a 3D interface that help in analyzing data more deeply and very helpful in prediction. For the classification of data, Weka 3.8 is used in the experiment [13], [19]–[21]. The data set used in this experiment is large. So, when uploading in Weka with default java memory setting than found the Java Virtual Heap size error which is solved by increasing the heap size, using the command prompt and change directory to Weka-3-8 and use the command: Weka-3-8-3> Java –Xmx5120m –jar weka.jar that increased the heap size up to 5GB, which make the loading of dataset easier.

C. Classifier

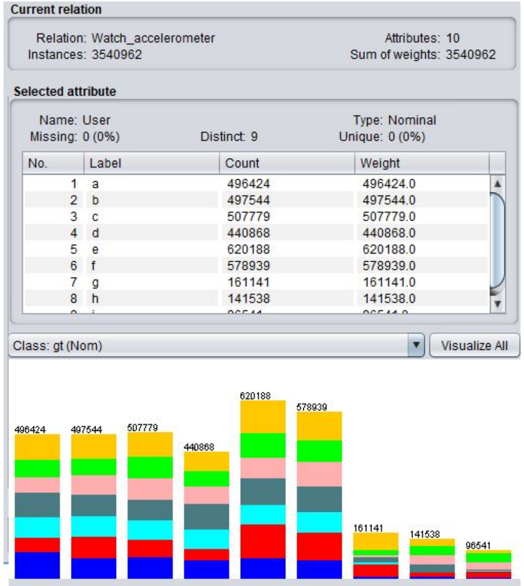
In this article, we used seven different built-in classifiers in Weka 3.8 to classify the seven activities (biking, walking, sitting, and standing, stair up, stair down, and null). The seven classifiers used are Decision Stump, J48, Bayes Net, Naive Bayes, Naive Bayes Multinomial Text, Random Forest, and Logit boost. The input data size is not fixed because Weka handles the huge amount of data to be processed and convert into. arff format but the machine Java virtual heap memory size can affect the loading of data in Weka. In this experiment, a huge amount of data, in which each file size is greater than 1GB of size having 6,746,393 instances. By increasing the Java Heap size, it becomes easier to execute the huge size of data in Weka.

The data is loaded in the preprocess tab available in Weka. The Preprocess screen describes the data related to the watch accelerometer and watches gyroscope files. Both files have different instances like the watch accelerometer has 3,540,962 instances and the watch gyroscope has 3,205,431 instances. After loading the. arff format in Weka. The screen shows the index attribute is selected by default and the dataset has been checked for having no missing values, no unique values. In Preprocess, a histogram shows how often an individual 10 selected values of classes like ground truth of nominal data set gt (nom) arises for the individual value of users are shown in Fig. 2.

TABLE I. Information about Dataset

Sensors	2	Accelerometer, Gyroscope
Smartwatch's	4	2 Samsung Galaxy Gears 2 LG watches
Users	9	a, b, c, d, e, f, g, h, i
Activities	7	Biking, Sitting, Standing, Walking, Stair up, Stair down, null
Attributes	10	Index, Arrival Time, Creation Time, X, Y, Z, User, Model, Device, Ground Truth

Fig. 1. Samsung Galaxy Gear & LG Smartwatch [17] [18].



Current relation

Relation: Watch_accelerometer
Instances: 3540962
Attributes: 10
Sum of weights: 3540962

Selected attribute

Name: User
Missing: 0 (0%)
Distinct: 9
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	a	496424	496424.0
2	b	497544	497544.0
3	c	507779	507779.0
4	d	440868	440868.0
5	e	620188	620188.0
6	f	578939	578939.0
7	g	161141	161141.0
8	h	141538	141538.0

Class: gt (Nom)

Fig. 2. Weka –Preprocess – Colors In Histogram Define the Users as "blue – Standing, Red – Null, Teal – Sitting, Steel – Walking, Light Pink – Stair Up, Green – Stair Down, Yellow – Bike".

Now onwards, after checking the dataset for no noisy data or missing values in it. The second step is to perform the classification of data. To apply classification on the dataset, the seven available classifiers in Weka such as Decision Stump, J48, Bayes Net, Naive Bayes, Naive Bayes Multinomial Text, Random Forest, and Logit boost are used. A decision stump is a decision tree, which uses only a single attribute for splitting.

Decision stump illustrates the learners with a high bias and low variance that's why often used as a weak learner. Decision stumps perform notably well on some commonly used benchmark datasets from the UCI repository [22][23].

The **J48 algorithm** is used to classify different applications and perform accurate results of the classification. J48 algorithm is one of the best machine learning algorithms to examine the data categorically and continuously [24].

A **Bayesian network** is a form of a directed graphical model for representing multivariate probability distributions [25][22].

Naïve Bayes is a simple learning algorithm that utilizes the Bayes rule together with a strong assumption that the attributes are conditionally independent, given the class [22].

Multinomial naive Bayes for text data. Operates directly (and only) on string attributes other. types of input attributes are accepted but ignored during training and classification [26].

Random Forests is an ensemble learning technique. It is a hybrid of the Bagging algorithm and the random subspace method and uses decision trees as the base classifier. Each tree is constructed from a bootstrap sample from the original dataset. The *"n"* is suggested to be $log_2 (N + 1)$, where N is the size of the whole feature set. Random forest classifier handles the missing values and maintains the missing data. Random forest handles the large dataset with higher dimensionality. In this research, the handling of the large dataset is the best fit with Random forest and the below results prove that also [3][2][22][23][27][28].

In machine learning and computational learning theory, **LogitBoost** is a boosting algorithm. The original paper casts the AdaBoost algorithm into a statistical framework. Specifically, if one considers AdaBoost as a generalized additive model and then applies the cost function of logistic regression, one can derive the LogitBoost algorithm [29].

The analysis is performed on both accelerometer and gyroscope datasets with the same seven classifiers. The Classify tab in Weka shows the Test Option, Result List, and Classifier output screen. In the Test Option, the percentage split option is chosen based on 66% training data and 34% testing data. Table II shows the splitting percentage of data.

TABLE II. DATA SPLITTING PERCENTAGE

Training Data (66%)		**Testing Data (34%)**	
Accelerometer	**Gyroscope**	**Accelerometer**	**Gyroscope**
2337035	2115584	1203927	1089847

IV. PERFORMANCE EVALUATION

The experiment is based on the performance of the following classifiers available in Weka software: Decision Stump, J48, Bayes Net, Naive Bayes, Naive Bayes Multinomial Text, Random Forest, and Logit boost. The classification is trained and tested using nine users and 10 attributes. The experiment is based on two parts: one is the time performance of the classifier and the other is the accuracy of the classifier.

A. Time Performance

The time performance of the classifier is shown in Table III on both the accelerometer and gyroscope dataset. In the time performance, the time is taken by the model to train itself, and onwards time is taken to test the model based on splitting the data as shown in Table II. Lesser the time is taken by the classifier justify that it doesn't mean that classification is done accurately. As in Table III, the Decision Stump shows that the time taken to test the model is 1.66s on the accelerometer and 1.13s on the gyroscope sensor. On the second, the J48 shows the time taken to test the model is 2.54s on the accelerometer and 1.72s on the gyroscope sensor. According to this result, the decision stump takes a fewer second to test the result but these results are accurate or not will be proof in Accuracy Rate shown in Table IV.

TABLE III. TIME PERFORMANCE

Time is taken to	**build the model**		**test the model**	
	Accelerometer	**Gyroscope**	**Accelerometer**	**Gyroscope**
Decision Stump	13.39s	12.94s	1.66s	1.13s
J48	247.41s	203.11s	2.54s	1.72s
Bayes Net	78.02s	88.25s	7.4s	10.36s
Naive Bayes	4.74s	4.63s	15.7s	13.4s
Naive Bayes Multinomial Text	0.23s	0.24s	5.08s	4.91s
Random Forest	1912.08s	1713.5s	39.52s	40.27s
Logit Boost	552.25s	498.24s	3.07s	2.24s

TABLE IV. ACCURACY SUMMARY OF THE CLASSIFIERS

Classifiers	**Accelerometer**	**Gyroscope**
Decision Stump	26.27%	27.13%
J48	99.98%	99.98%
Bayes Net	99.99%	99.98%
Naive Bayes	53.21%	48.58%
Naive Bayes Multinomial Text	17.90%	16.28%
Random Forest	99.99%	99.99%
Logit Boost	72.89%	74.39 %

B. Accuracy Rate

The classification accuracy is calculated as shown in equation (1). This value is the ratio of the correctly classified sample number to the total sample number.

$$\text{Accuracy (CA)} = \frac{\text{TP+ TN}}{\text{TP} + FP\text{+FN +TN}} * 100\% \qquad (1)$$

Where TP is the number of true positives, TN defines the number of true negatives, FP is the number of false positives and FN is the number of false negatives [3].

The accuracy summary of all classifiers shown in Table IV. Among all classifiers the Random Forest, Bayes Net, and J48 perform were well. As in Table III, the time performance of Decision Stump and J48 is fast to execute but accuracy to classify the data is not efficient to be used for a huge amount of data. However, our experiment shows that the Random Forest took much time to build the model and test the model but the result shown are accurately classified. The Random Forest [14] [7] [4] verified the high accuracy rate of 99.99% on both datasets of accelerometer and gyroscope. Below Fig. 3 shows the graphical representation of the accuracy rate of the used classifiers.

In Tables V and VI, the confusion matrix is based on the Random Forest classifier. The matrix shows the accurately classified instance for accelerometer and gyroscope sensors. It gives an overview of all activities like standing, null, sitting, walking, stair up, stair down, and on a bike.

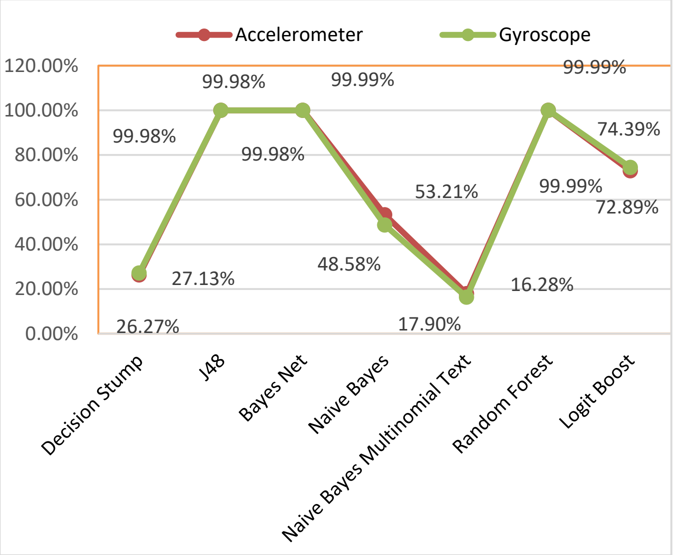


Fig. 3. Shows the accurate of the Classifiers.

TABLE V. CONFUSION MATRIX FOR ACCELEROMETER SENSOR

Accelerometer							
a	b	c	d	e	f	g	
153126	1	0	0	0	0	0	a
2	177357	2	3	2	0	6	b
0	5	144415	0	0	0	0	c
0	1	0	186628	0	0	0	d
0	5	0	0	160844	30	0	e
0	4	0	0	36	165873	0	f
0	2	0	0	0	0	215585	g

TABLE VI. CONFUSION MATRIX FOR GYROSCOPE SENSOR

Gyroscope							
a	b	c	d	e	f	g	
146222	4	0	0	0	0	0	a
2	159839	4	3	1	1	6	b
1	3	142790	0	0	0	0	c
0	6	0	166121	0	0	0	d
0	2	0	0	151856	25	0	e
0	0	0	0	44	145395	0	f
0	2	0	0	0	0	177520	g

*a = stand, b = null, c = sit, d = walk, e = stairs up, f = stairs down, g = on bike

V. CONCLUSION

In this study, the features obtained after applying classification to the data set generated by the accelerometer and gyroscope sensor. The smartwatches are classified using machine learning methods. In this study, where seven different daily human activities are classified, the most successful result is obtained from the random forest method. The Random Forest is easy to understand and handle the large dataset. The result shows that Random Forest classified 99.99% correct classification on both datasets using the Weka toolkit.

The limitation of this work is the hardware used and Weka toolkit package support on 6GB RAM and 5GB Java. The used hardware effects time performance of many classifiers also restrict some move to use because of processing took longer time. Furthermore, if the resources are updated then the time performance evaluation can become better and some classifiers can be used.

In the future, the updated resources and predictions from this analysis work in the human recognition system. The result of this study can help in the recognition of activities to save energy consumption. Applications of detection of human activities can support people in healthy life subject. It can help detect and prevent dangerous actions such as falling and disappearing of older people and young children, or actions which are not good for the health of a person. Furthermore, the dataset can be changed to unsupervised data and use clustering and association algorithms to perform prediction and making the human recognition support system better.

REFERENCES

[1] F. Shahmohammadi, A. Hosseini, C. E. King, and M. Sarrafzadeh, “Smartwatch Based Activity Recognition Using Active Learning,” in Proceedings - 2017 IEEE 2nd International Conference on Connected Health: Applications, Systems, and Engineering Technologies, CHASE 2017, 2017, pp. 321–329.

[2] P. Fulfillment et al., “Developing Information and Communication Technology ’ S (ICT) Acceptance Model in the Context of Social Environment and Measuring the Performance using Fuzzy Models,” 2017.

[3] S. Balli, E. A. Sağbaş, and M. Peker, “Human activity recognition from smartwatch sensor data using a hybrid of principal component analysis and random forest algorithm,” Meas. Control (United Kingdom), vol. 52, no. 1–2, pp. 37–45, 2019.

[4] D. D. Adrian Fernandez, “Accelerometer Sensor,” 2013. [Online]. Available: https://www.sciencedirect.com/topics/engineering/accelerometer-sensor.

[5] 2019 Goleman et al., Chapter 4, Using the Accelerometer, vol. 53, no. 9. O’Reilly Media, Inc., 2019.

[6] M. Barela, Make: Getting Started with Circuit Playground Express. 2018.

[7] H. M. Ali and A. M. Muslim, “Human Activity Recognition Using Smartphone and Smartwatch,” Int. J. Comput. Eng. Res. Trends, vol. 3, no. 10, pp. 2349–7084, 2016.

[8] R. J. Oskouei, Z. MousaviLou, Z. Bakhtiari, and K. B. Jalbani, “IoT-Based Healthcare Support System for Alzheimer’s Patients,” Wirel. Commun. Mob. Comput., vol. 2020, pp. 1–15, Oct. 2020.

[9] A. Stisen et al., “Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition,” SenSys 2015 - Proc. 13th ACM Conf. Embed. Networked Sens. Syst., pp. 127–140, 2015.

[10] M. C. Kwon, H. You, J. Kim, and S. Choi, “Classification of Various Daily Activities using Convolution Neural Network and Smartwatch,” in Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018, 2019, pp. 4948–4951.

[11] Z. Zhuang and Y. Xue, “Sport-related human activity detection and recognition using a smartwatch,” Sensors (Switzerland), vol. 19, no. 22, pp. 1–21, 2019.

[12] A. Bayat, M. Pomplun, and D. A. Tran, “A study on human activity recognition using accelerometer data from smartphones,” Procedia Comput. Sci., vol. 34, no. December, pp. 450–457, 2014.

[13] S. Srivastava, “Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering, and Association Rule Mining,” Int. J. Comput. Appl., vol. 88, no. 10, pp. 26–29, 2014.

[14] L. M. S. Morillo, L. Gonzalez-Abril, J. A. O. Ramirez, and M. A. A. De La Concepcion, “Low energy physical activity recognition system on smartphones,” Sensors (Switzerland), vol. 15, no. 3, pp. 5163–5196, 2015.

[15] G. M. Weiss, K. Yoneda, and T. Hayajneh, “Smartphone and Smartwatch-Based Biometrics Using Activities of Daily Living,” IEEE Access, vol. 7, pp. 133190–133202, 2019.

[16] G. M. Weiss, “WISDM Smartphone and Smartwatch Activity and Biometrics Dataset,” UCI Mach. Learn. Repos. WISDM Smartphone Smartwatch Act. Biometrics Dataset Data Set, vol. 7, pp. 133190–133202, 2019.

[17] Samsung, “Samsung Galaxy Gear,” Samsung. [Online]. Available: https://www.samsung.com/in/support/mobile-devices/is-samsung-galaxy-gear-waterproof/. [Accessed: 18-Dec-2020].

[18] LG, “LG Smartwatch.” [Online]. Available: https://www.lg.com/us/smart-watches/lg-W100-lg-watch.

[19] E. Frank, M. A. Hall, and I. H. Witten, “The WEKA workbench,” Data Min., pp. 553–571, 2017.

[20] “Weka 3: Data Mining Software in Java.” [Online]. Available: https://www.cs.waikato.ac.nz/~ml/weka/.

[21] C. J. P. Ian H. Witten, Eibe Frank, Mark A. Hall, Data Mining Practical Machine Learning Tools and Techniques, 4th ed.

[22] K. M. Ting, Precision and Recall BT - Encyclopedia of Machine Learning. 2010.

[23] P. Morgan, Introduction to Data Science: Essential Concepts, vol. 53, no. 9. AI Sciences LLC, 2013.

[24] N. S. anaN and V. G. thri, “Performance and Classification Evaluation of J48 Algorithm and Kendall’s Based J48 Algorithm (KNJ48),” Int. J. Comput. Trends Technol., vol. 59, no. 2, pp. 73–80, 2018.

[25] S. M. Aqil Burney and J. Naseem, “Decision Making in Uncertainty : A Bayesian Network for Plant Disease Diagnoses,” Int. J. Comput. Sci. Inf. Secur., no. April 2018.

[26] Weka, “Class NaiveBayesMultinomialText.” [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/classifiers/bayes/NaiveBayesMultinomialText.html. [Accessed: 28-Nov-2020].

[27] L. Breiman, “Random Forests,” 2001.

[28] Scott Hartshorn, Machine Learning With Random Forests And Decision Trees A Visual Guide For Beginners. 1393.

[29] J. Friedman, T. Hastie, and R. Tibshirani, “Additive Logistic Regression,” The Annals of Statistics, vol. 28, no. 2. pp. 337–374, 2000.

Performance Analysis of Fermat Factorization Algorithms

Hazem M. Bahig[1]*, Mohammed A. Mahdi[2], Khaled A. Alutaibi[3], Amer AlGhadhban[4], Hatem M. Bahig[5]
Computer Science and Information Department, College of Computer Science and Engineering, University of Ha'il, Ha'il, KSA[1,2]
Computer Science Division, Mathematics Department, Faculty of Science, Ain Shams University, Cairo, Egypt[1,5]
Computer Engineering Department, College of Computer Science and Engineering, University of Ha'il, Ha'il, KSA[3]
Electrical Engineering Department, College of Engineering, University of Ha'il, Ha'il, KSA[5]

***Abstract*—The Rivest-Shamir-Adleman (RSA) cryptosystem is one of the strong encryption approaches currently being used for secure data transmission over an insecure channel. The difficulty encountered in breaking RSA derives from the difficulty in finding a polynomial time for integer factorization. In integer factorization for RSA, given an odd composite number *n*, the goal is to find two prime numbers *p* and *q* such that *n* = *p q*. In this paper, we study several integer factorization algorithms that are based on Fermat's strategy, and do the following: First, we classify these algorithms into three groups: Fermat, Fermat with sieving, and Fermat without perfect square. Second, we conduct extensive experimental studies on nine different integer factorization algorithms and measure the performance of each algorithm based on two parameters: the number of bits for the odd composite number *n*, and the number of bits for the difference between two prime factors, *p* and *q*. The results obtained by the algorithms when applied to five different data sets for each factor reveal that the algorithm that showed the best performance is the algorithms based on (1) the sieving of odd and even numbers strategy, and (2) Euler's theorem with percentage of improvement of 44% and 36%, respectively compared to the original Fermat factorization algorithm. Finally, the future directions of research and development are presented.**

Keywords*—Integer factorization; Fermat's algorithm; RSA; factorization with sieving; perfect square

I. Introduction

The Rivest-Shamir-Adleman (RSA) cryptosystem is one of the most famous and secure cryptosystems currently available. It was designed to encrypt plain text into cipher text in as strong a manner as possible. The RSA system is a type of public-key cryptosystem that is based on two different keys: a public key that is used for encryption and a private key that is used for decryption.

The main steps in the RSA cryptosystem are as follows [1,2]:

1) Generate two random distinct prime numbers of large and equal size, p and q, and then construct an odd composite number *n* = *p q*.

2) Calculate the Euler function $\Phi(n) = (p-1)(q-1)$.

3) For the encryption procedure, choose the exponent number e that is greater than 1 and less than $\Phi(n)$ such that $\gcd(e, \Phi(n)) = 1$. Then apply the modular exponentiation formula on the message m to generate a secret message c as follows:

$$c = m^e \bmod n$$

4) For the decryption procedure, find the integer d that is greater than 1 and less than $\Phi(n)$ such that $e\, d \bmod \Phi(n) = 1$. Then apply the modular exponentiation formula on the secret message *c* to generate a message *m* as follows:

$$m = c^d \bmod n$$

The RSA cryptosystem includes two mathematical operations that are opposite to each other. The first operation is multiplication, which is easy to compute. The running time to compute the product of two numbers is $O(b^2)$ in the worst case, where *b* is the size of each number. This type of operation is important for computing the modular exponentiation [3, 4] to reduce the computation time of the exponentiation. The second operation is a process that involves finding two prime factors *p* and *q* from an odd composite number *n.* This process is called integer factorization [5]. If we can factor *n* to *p* and *q*, then we can compute $\Phi(n)$ and then d. Consequently, the encrypted message *c* can be decrypted. Hence, the integer factorization problem is important in cryptography. Therefore, solving this problem in an efficient timeframe leads to breaking the RSA. In other side, the difficulty in finding a polynomial time for the factorization leads to difficulty in breaking the RSA cryptosystem [6, 7, 8, 9].

Moreover, the integer factorization problem is important from the point of view of complexity theory. Until now, the integer factorization problem has not been considered to belong to the class of P problems. Also, there is no proof that the integer factorization problem belongs or does not belong to the class of NP-complete problems. From a review of the literature, it seems that the best time complexity for factoring an odd composite number is

$$exp\left(\sqrt[3]{64/9} + o(1)(\ln n)^{1/3}\,(\ln\ln n)^{2/3}\right)$$

using the general number field sieve (GNFS) algorithm [1, 10].

A large number of algorithms have been proposed in order to attempt to factor an odd composite number. These algorithms can be categorized as either general or special-purpose algorithms. The general-purpose group contains

*Corresponding Author

integer factorization algorithms that have a running time that depends on the size of an odd composite number only. This group includes integer factorization algorithms that are based on various strategies, such as continued fraction factorization, Shanks's square forms factorization, Dixon's algorithm, the quadratic sieve algorithm and the GNFS algorithm [1, 2, 10].

On the other hand, the special-purpose group contains integer factorization algorithms that have a running time that depends on the size of an odd composite number and its properties. For example, the trial division method is an efficient algorithm for factorization when an odd composite contains a small prime factor. Besides trial division, the special-purpose group contains various other techniques, such as Fermat factorization, wheel factorization, Pollard's p-1, Euler factorization, and the Lenstra elliptic curve [1, 2, 10, 11].

In this paper, we are interested in the special-purpose group because the aim is to study the performance of algorithms that are based on Fermat's factorization concept. Fermat proposed a factorization algorithm that is based on representing the odd composite number as the difference between two squares. The main advantage of Fermat's factorization technique is that it is able to factor an odd composite number, n, in a very fast time, i.e., almost instantaneously, when the difference between two factors is $\Delta = \sqrt[4]{n}$ [6, 12]. This means that if the size of n is b bits and the difference between two factors is $\Delta = \sqrt[4]{n}$, then the following are true: (1) The two prime factors have the same size, i.e., each prime factor has size $b/2$; (2) The number of common bits between the two prime factors is $b/4$, and these bits should be the most significant bits. These two conditions are known as the domain of the efficiency of Fermat's algorithm, (DEF).

Many algorithms have been proposed that are based on Fermat's factorization concept [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31]. The goal of these algorithms is to improve the running time of the original Fermat algorithm in finding prime factors. Two categories of factor have an effect on the efficiency of these algorithms. The first category is related to the properties of the input, which includes the size of the odd composite number, *b*, and the difference between two factors, Δ. The second category is related to the natural of the algorithm itself such as the search strategy it uses to find the solution and the number of high-cost operations included in the algorithm.

In general, the improved Fermat algorithms can be classified into two classes. The first class contains algorithms based on the concept of an estimated prime factor and uses different techniques such as continued fraction method [28] or considering *n* as a special form $6k \pm 1$, where *k* is any integer [23]. However, the techniques in this class cannot factor some odd composite numbers, so they cannot be considered as general methods for Fermat factorization. The second class contains algorithms [11, 14, 15, 17, 18, 19, 20, 21, 22, 24, 25, 26, 27, 29] that can be applied to any odd composite number and are based on (1) replacing the high-cost operation, i.e., the perfect square in Fermat's method, with a low-cost operation or on (2) reducing the space searched to find the solution. It should also be noted that there is another strategy [13, 30] that falls outside the scope of our research, which involves speeding up the running time of Fermat's algorithm that is based on a different platform such as high-performance computing [13, 33].

In this paper, we are interested in the integer factorization algorithms that are belong to the second class. From our analysis of these techniques, we made the following observations:

1) The experimental studies for most of these algorithms were implemented when the size of an odd composite was less than 64 bits [15, 20, 21, 26, 32], for example, n = 84449 × 21121 =1783647329. This number of bits is small compared to that required in cryptography. Also, the time consumed for an operation increases with increase the size of data, especially for high-cost operations.

2) Many of the experimental studies for some of these algorithms were implemented when the difference between the two factors did not belong to the DEF. This means that any comparison between Fermat's algorithm and these algorithms is not realistic because it has been proved that the Fermat factorization method is not efficient outside the DEF. For example,n=19710741 × 531349691=1047329636821139813 [20].

3) The efficiency of some of these algorithms was measured based on a few data or on some examples, rather than on different values for *b* and Δ, see for example, [22, 25]. This means that there is no exhaustive study that compares two or more integer factorization algorithms that are based on Fermat factorization concept by applying them to different data distributions in the DEF.

4) A few steps in some of these algorithms required some optimization due to the cost of the operation to manipulate a large data size.

Consequently, we are interested in undertaking an experimental study on most of the integer factorization algorithms that are based on Fermat's concept in order to answer the following:

Q1) Which one of the Fermat factorization algorithms is efficient for a large size of *n* and a small value of Δ?

Q2) What is the effect of increasing the value of Δ with a fixed size of *n* for each of the studied algorithms?

Q3) Many integer factorization algorithms have the same number of iterations, theoretically, but which one is the fastest over different data distributions?

To the best of our knowledge, there is no sufficient comparative study for Fermat factorization algorithms, especially with regard to the effect of the use of factors *b* and Δ on performance. Also, our study compares the performance of nine different integer factorization algorithms in order to determine which has the fastest running time.

The remainder of the paper is structured as follows: In Section II, we provide the methodology used to verify the objectives of this study. In Section III, we provide an overview of the different integer factorization algorithms that are based on Fermat factorization, including the pseudocode of each algorithm. Additionally, we classify these algorithms based on the techniques used into three groups. In Section IV, we describe the experimental study undertaken to compare and measure the performance of different integer factorization algorithms. Also, we present an analysis of the results produced by the experimental study. Finally, in Section V, we draw some conclusions from this work and highlight open questions that remain to be answered in future studies.

II. Methodology

To measure the performance of the integer factorization algorithms that are based on Fermat's strategy, we followed a methodology that consisted of five stages:

1) Determine the different strategies that need to be used to factor an odd composite integer into two prime factors according to Fermat's concept;

2) Determine the language and packages to use to verify the goal of the study;

3) Optimize the implementation of the selected integer factorization algorithms based on the platform used in the experimental studies;

4) Generate a dataset to use to measure the performance of the selected algorithms, especially when applied to large data sizes; and

5) Measure and analyze the performance of the selected algorithms.

Here we discuss, briefly, each of the above stages. In the first stage, we studied the different algorithms that use Fermat's strategy to find the two prime factors for an odd composite number. We classified these algorithms into groups based on the concept used in the algorithms. For each algorithm, we identified the main idea, the pseudocode and the expected number of iterations to find the solution. The details of this stage are covered in Section III.

In the second stage, we selected the language and package to use in our study, namely, C++ language and the GNU Multiple Precision (GMP) arithmetic library [34]. We decided to use C++ language because the performance of this language is fast compared to other languages such as Java. In other side, to execute any operation in the RSA system requires a number of size 1024 bits. However, the size of the integer type in C++ language does not support this objective because it is limited to 64 bits. Therefore, we decided to use the GMP library because it is designed to support applications such as cryptography and computational algebra that involve large-sized numbers. Furthermore, the library has the following advantages: (1) It contains a large number of functions to manipulate integers and other types; (2) the functions in the library are fast compared to those of other tools; and (3) there is no limitation to the size of number, so we can manipulate a number that is greater than 1024 bits in size.

It should be noted that, in our implementation, we used only the data type mpz_t that is used to manipulate GMP integers. The library contains a group of functions to manipulate GMP integers, such as (1) initializing and assigning GMP integers, (2) integer arithmetic and division, (3) integer roots, and (4) integer comparisons.

In the third stage, we focused on optimizing each algorithm, if required, in order to run the algorithm in a fast way. The reasons for doing this were as follows: (1) The cost of the operation for a large integer size is significantly different than that for small integers of less than 64 bits for the same operation [35, 36]; and (2) for some algorithms, we needed to rewrite a few of the statements to increase the performance of an algorithm. The details of this stage are provided in Section III.

In the fourth stage, we employed a method to generate an odd composite number consisting of a product of two prime numbers. In this method, the generation of the two prime factors is based on the two factors of the DEF [6, 12]. The first factor is the size of the odd composite number n. Suppose that the size of n is b, where b is the number of bits. Therefore, when we generate an odd composite number n of size b bits, we first generate two prime numbers p and q, each of size $b/2$, and then we multiply both of them, i.e., $n = p\ q$. The second factor is the difference between the prime factors p and q, Δ. This factor is important because the running time of Fermat's strategy increases with an increase in the value of Δ.

The generation of one data (an odd composite number), GD, consists of the following steps: The first step is to determine the number of bits for n, b, and the number of bits for the difference between two factors, Δ. The second step is to generate a prime number of size $b/2$, say p. The third step is to generate a random number of size Δ and then add it to the first prime number, say x. The fourth and final step involves generating the second prime number, say q, greater than x such that the size of the difference between p and q is Δ.

For more accuracy in measuring the performance of the different algorithms, we repeated the previous steps for GD by adopting the following procedure: First, we fixed the value of b and Δ. Second, we applied the steps for generating two prime factors as in GD, i.e., from the second step to the fourth step in GD, t times, where t represents the number of different instances that have the same value of b and Δ. Third, we repeated the execution of the first and second steps of the procedure with increasing values of Δ in increments of 5 bits until $\Delta + 20$. The reason for setting a maximum value of Δ is that the running time of all the integer factorization algorithms based on Fermat's strategy increases rapidly with an increase in Δ. Fourth, we increased the value of b and then repeated all the previous steps.

The last important task in the data generation stage was to verify that the generated data were correct as follows:

1) Each prime factor, p, should be verified such that $2^{(b/2)-1} < p < 2^{(b/2)}$.

2) The difference between the two factors should be verified such that $2^{\Delta-1} \leq |p-q| < 2^{\Delta}$.

3) The prime factors used for the fixed values of b and Δ should be as different as possible.

The fifth and final step in the methodology involved measuring the performance of the integer factorization algorithms that are based on Fermat's strategy. The performance of these algorithms was mainly measured by computing the execution time. Hence, for fixed values of b and Δ, the running time for the algorithm A, $T_A(b,\Delta)$, computed using the following formula:

$$T_A(b,\Delta) = \frac{1}{t}\sum_{i=1}^{t} T_A(b,\Delta)_i$$

where $T_A(b,\Delta)_i$ is the running time of the algorithm A for the instance number i using input data b and Δ. Note that two instances, $(b,\Delta)_i$ and $(b,\Delta)_j$, are different if the odd composite number, n_i, of size b for the instance i is different than the odd composite number, n_j, of size b for the instance j. Additionally, with respect to the issue of memory consumption, all the algorithms required a constant number of auxiliary variables, so there was no need to measure this factor experimentally.

III. Classification of Fermat Algorithms

In this section, we provide an overview of the different algorithms for integer factorization that are based on Fermat's concept. For each algorithm, we discuss the main idea and steps of the algorithm, and then we give the pseudocode of the algorithm.

Without loss of generality, for all algorithms, we assume that the integer number n is odd and is a product of two prime numbers p and q, where $p > q$. The main idea of Fermat's algorithm is that the integer number n can be expressed as a difference between two square numbers, x and y. Formally, the odd integer number n can be written as follows:

$$n = x^2 - y^2 \tag{1}$$

Also, the relation between the two prime factors and the two square numbers is as follows:

$$p = x + y \text{ and } \quad q = x - y \tag{2}$$

Different strategies have been proposed to factorize an odd composite number into two prime numbers based on Fermat's concept. All the algorithms start the search with an initial value of x and try to find the value of y such that $y^2 = x^2 - n$. So, the main issues in Fermat's strategy are (1) how to reduce the search space of x and (2) how to reduce the cost of the perfect squaring operation, where a simple test for a perfect square for x includes two operations: calculating the square root for x, say r, and testing whether the value of r is an integer or not.

We can classify the algorithms of integer factorization that are based of Fermat's strategy into three main groups. The first group employs a direct approach which starts from the minimum value of x and uses a perfect square operation. We named this group the Fermat factorization group because it is based on directly applying the concept proposed by Fermat. The second group is based on sieving or pruning some elements in the search space, so the algorithm does not apply the squaring operation or/and perfect squaring operation on those elements. We named this group the Fermat with sieving group. The third group is based on removing the main operation of the Fermat factorization algorithm which is the perfect squaring operation, so we named this group the Fermat without perfect squaring group. The main ideas and steps of these three groups of algorithms are outlined in the following subsections.

A. The Fermat Factorization Group

The Fermat factorization group contains many algorithms which are based mainly on the perfect squaring operation [14, 15, 17]. The first and main algorithm in this group is based on rewriting Eq. (1) as $y^2 = x^2 - n$ and starts by assuming that the value of x is $\lfloor\sqrt{n}\rfloor + 1$. Then the algorithm tests whether the value of $x^2 - n$ is a perfect square. If the value of y is an integer and equal to $\sqrt{x^2 - n}$, the search is terminated. When $x^2 - n$ is not a perfect square, the algorithm increases the value of x by 1 and follows the same procedure until it finds the value of y.

All the steps of the algorithm are shown in Algorithm FF. The algorithm is very straightforward and contains simple operations, except for the perfect square operation. The running time of the algorithm is based on two factors. The first factor is the cost of the perfect square operation and the second is the number of iterations for the While-loop which is equal to $x - \left(\lfloor\sqrt{n}\rfloor + 1\right)$ in the worst case, where x satisfies that the term $x^2 - n$ is a perfect square and equals $(p+q)/2$.

Algorithm FF (Fermat's Factorization)

Input: n is a positive odd number.

Output: p and q are two prime numbers such that $n = p\,q$.

```
Begin
    $x = \lfloor\sqrt{n}\rfloor + 1$
    $y = x^2 - n$
    While ($y$ is not a perfect square) do
        $x = x + 1$
        $y = x^2 - n$
    End while
    $p = x + \sqrt{y}$
    $q = x - \sqrt{y}$
End.
```

Remarks:

1) Many modified algorithms [14, 15, 17] have been proposed to improve the FF algorithm while retaining the two operations, squaring and perfect squaring. The modifications are based on rewriting the Fermat factorization formula and then searching for the solution. For example, in [17], the formula is rewritten as $z^2 = \left(x + \lceil\sqrt{n}\rceil y\right)^2 - ny^2$, with a small x and y, and the goal is to find the solution (x, y, z). In [15], the algorithm is modified in order to achieve the goal of finding a solution for $z^2 = \left(\lceil\sqrt{n\,i}\rceil\right)^2 - n\,i$, where i starts with a value of 1. The modified algorithm, FF1, contains four operations before testing the perfect square. These operations are: multiplying n with i, squaring the root of (n i), say s, squaring s, and calculating the modulus of n. This means that, in general, the modified algorithm contains more operations than the FF algorithm. Therefore, we neglected these modifications in our general comparison of the different types of integer factorization algorithms (see Section IV).

2) In order to optimize the code of the FF algorithm we decided to do the following: (*i*) To compute x^2, we multiplied x with itself to get better performance instead of using the predefined function power for the exponent 2; (*ii*) in the case of the perfect square operation, we used the predefined function in GMP because we considered that this would be better than computing the square root of the number and then testing whether the results are integers or not.

Algorithm FF1 (Modified Fermat's Factorization)

Input: *n* is a positive odd number.

Output: p and q are two prime numbers such that $n = p\,q$.

```
Begin
    $i = 1$
    found =false
    While ($found \neq$ true) do
        $s = \lceil\sqrt{n\,i}\rceil$
        $m = s^2 \mod n$
        If IsSquare($m$) then
            $t = \sqrt{m}$
            found =true
            return gcd($n$,$s$-$t$)
        End if
        $i = i + 1$
    End while
End.
```

B. The Fermat with Sieving Group

The sieving strategy is a method that is used to remove the impossible solutions so that the algorithm does not consider them during the search process. The algorithms that apply this strategy for Fermat factorization can be classified into two classes of techniques.

The first class of techniques ignores the perfect squaring operation in some cases. This means that before testing whether integer *y* is a perfect square or not, the technique tests whether y satisfies a certain condition. If integer y meets this condition, the technique does not test whether y is a perfect square and goes instead to the next value of *x*. Otherwise, the technique tests whether *y* is a perfect square or not.

The second class of techniques ignores the squaring operation and consequently the perfect square operation. This means that before calculating x^2, the strategy tests whether *x* satisfies a certain condition. If integer *x* meets this condition, the technique ignores all the subsequent steps, i.e., squaring, subtraction, and perfect squaring, and goes to the next value of *x*. Otherwise, the technique tests whether y is a perfect square or not.

1) Class 1: Ignoring the perfect square

a) Sieving with modulus: One of the techniques used in sieving is the modulus operation, mod. The idea behind using the modulus arithmetic operation is to exclude all integers, x, that are definitely not perfect squares before applying the perfect squaring operation [22, 24].

TABLE I. VALUES OF X AND X^2

x	x^2	x	x^2
$x_{l-1}x_{l-1}\dots x_1 0$	$x'_{2l-1}x'_{l-1}\dots x'_1\mathbf{0}$	$x_{l-1}x_{l-1}\dots x_1 5$	$x'_{2l-1}x'_{l-1}\dots x'_1\mathbf{5}$
$x_{l-1}x_{l-1}\dots x_1 1$	$x'_{2l-1}x'_{l-1}\dots x'_1\mathbf{1}$	$x_{l-1}x_{l-1}\dots x_1 6$	$x'_{2l-1}x'_{l-1}\dots x'_1\mathbf{6}$
$x_{l-1}x_{l-1}\dots x_1 2$	$x'_{2l-1}x'_{l-1}\dots x'_1\mathbf{4}$	$x_{l-1}x_{l-1}\dots x_1 7$	$x'_{2l-1}x'_{l-1}\dots x'_1\mathbf{9}$
$x_{l-1}x_{l-1}\dots x_1 3$	$x'_{2l-1}x'_{l-1}\dots x'_1\mathbf{9}$	$x_{l-1}x_{l-1}\dots x_1 8$	$x'_{2l-1}x'_{l-1}\dots x'_1\mathbf{4}$
$x_{l-1}x_{l-1}\dots x_1 4$	$x'_{2l-1}x'_{l-1}\dots x'_1\mathbf{6}$	$x_{l-1}x_{l-1}\dots x_1 9$	$x'_{2l-1}x'_{l-1}\dots x'_1\mathbf{1}$

We can apply this technique as follows: For any integer *x*, we can represent *x* in decimal form as $x_{l-1}x_{l-1}\dots x_1x_0$, where *l* represents the number of decimal digits in *x*. It is clear that the last, right-most, digit of *x* is either 0, 1, 2, 3, 4, 5, 6, 7, 8, or 9. The last digit for squaring *x* can be calculated by taking the modulus of 10 and is equal to 0, 1, 4, 5, 6, or 9, see the bold digit in Table I. On the other hand, no squaring number has a last digit of 2, 3, 7 or 8. Therefore, we compute $r = y$ mod 10 and if the value of r is 2, 3, 7, or 8, then there is no need to test whether y is a perfect square or not, and so we can go to the next value of *x* directly.

The complete steps of the algorithm is shown in Algorithm FM10 [22, 24]. The algorithm is similar to the FF algorithm but contains two extra statements. The first statement computes the modulus of 10 for the term $x^2 - n$, and the second tests the result of using the modulus, see line 7. The running time of the algorithm is similar to that of the FF algorithm, except (1) two extra statements, see lines 6 and 7, and (2) the algorithm uses fewer perfect square operations based on the truth value of the condition.

Algorithm FM10 (Fermat Sieving by Modulus 10)

Input: *n* is a positive odd number.

Output: p and q are two prime such that $n = p\,q$.

```
Begin
1.  $x = \lfloor\sqrt{n}\rfloor$
2.  found=false
3.  While (Not found) do
4.      $x = x + 1$
5.      $y = x^2 - n$
6.      $r = y$ mod 10
7.      If Not ($r = 2$ or $r = 3$ or $r = 7$ or $r = 8$) then
8.          If ($y$ is a perfect square) then
9.              found=True
10.         End if
11.     End if
12. End while
13. $p = x + \sqrt{y}$
14. $q = x - \sqrt{y}$
End.
```

Remarks:

1) The statement in line 7 can be rewritten as follows:

If $(r = 0$ or $r = 1$ or $r = 4$ or $r = 5$ or $r = 6$ or $r = 9)$ then

However, this statement contains six comparisons at most, whereas the statement in line 7 contains four comparisons at most, but the running time for two versions is almost similar.

2) We can use a modulus of 15, 20 or 30 instead of 10. To study the effect of changing the value of the modulus on the performance of the algorithm, we changed the modulus of 10 to a modulus of 20 and named this method FM20. Using this approach, the accepted cases to ignore the test for the perfect square occur when the remainder of (*r*=*y* mod 20) are 0, 1, 4, 5, 9, or 16. The steps of the FM20 algorithm are similar to those of FM10, except that line 7 is replaced with:

If Not $(r = 0$ or $r = 1$ or $r = 4$ or $r = 5$ or $r = 9$ or $r = 16)$ then

We studied the effect of this change experimentally, see Section IV.

b) Sieving with odd & even: In the FF algorithm, the values of x are $\lfloor\sqrt{n}\rfloor + 1$, $\lfloor\sqrt{n}\rfloor + 2$, $\lfloor\sqrt{n}\rfloor + 3$, $\lfloor\sqrt{n}\rfloor + 4$,…. This means that the values of *x* are odd and even numbers. Another sieving technique that can be applied in integer factorization algorithms is based on ignoring all the even (or all the odd) numbers of *x* if the integer *n* satisfies a certain condition. The idea behind using the even and odd property is based on the following rules [19, 29]:

1) Any odd integer n can be expressed as $n = 4\,k \pm 1, n \geq 3$.

2) For $n = 4\,k \pm 1$, we have two cases: (*i*) when $n = 4\,k + 1$, then *x* is odd and *y* is even, and (*ii*) when $n = 4\,k - 1$, then *x* is even and *y* is odd.

The algorithm consists of four main steps. The first step determines the form of n as either $4\,k + 1$ or $4\,k - 1$. The second step determines the type (even or odd) of *x* and *y*. The third step determines the start value of *x* in the case of whether *x* is odd or even. The fourth step applies the steps of FF algorithm with updating the value of *x* with 2.

The complete steps of the algorithm are shown in Algorithm FOE. To determine the formula of *n*, the algorithm computes the remainder of dividing *n* with 4 and then tests if the remainder is equal to 1 or not, see lines 2–3. If the remainder is equal to 1, this means that $n = 4\,k + 1$, otherwise, $n = 4\,k - 1$. Lines 4–5 and 8–9 determine the type, odd or even, and the start value of the integer *x*. The remainder of the algorithm is similar to the FF algorithm except when the value of $x^2 - n$ is not a perfect square, the algorithm updates the value of *x* by 2 instead of 1 because when *x* is even (or odd), the next even (or odd) number of *x* can be found by adding 2 to *x*. Hence, the number of iterations of the FOE algorithm is half that of the FF algorithm because the FOE algorithm updates the value of *x* by 2, whereas the FF algorithm updates the value of *x* by 1.

Algorithm FOE (Fermat Sieving using Odd & Even)

Input: *n* is a positive odd number.

Output: p and q are two prime numbers such that $n = p\,q$.

```
Begin
  1.  $x = \lfloor\sqrt{n}\rfloor + 1$
  2.  $r = n \bmod 4$
  3.  If $(r = 1)$ then
  4.      If ($x$ is even) then
  5.          $x = x + 1$
  6.      End if
  7.  Else
  8.      If ($x$ is odd) then
  9.          $x = x + 1$
 10.      End if
 11.  End if
 12.  $y = x^2 - n$
 13.  While ($y$ is not a perfect square) do
 14.      $x = x + 2$
 15.      $y = x^2 - n$
 16.  End while
 17.  $p = x + \sqrt{y}$
 18.  $q = x - \sqrt{y}$
End.
```

3) Class 2: Ignoring the Squaring: Another important improvement that has been made to Fermat's algorithm is the ignoring of some elements in the search space before squaring the value of *x*. Two algorithms have been proposed to achieve this goal. The two algorithms are based on analyzing the relation between the value of (*x* mod 10) and the value of (*n* mod *m*), where *m* may be 10, 15, 20, or 30.

For any integer *x*, the value of (*x* mod *m*) is either 0, 1, 2, …, *m*-2, or *m*-1. When *n* is odd and not divisible by 5 and *m* = 10, the value of (*n* mod 10) is 1, 3, 7, or 9, and when *m* = 20, the value of (*n* mod 20) is 1, 3, 7, 9, 11, 13, 15, 17, or 19. We ignored the value 5, because if (*n* mod 20)=5, then 5 is a factor of *n*.

TABLE II. DIFFERENT CASES FOR (X MOD 10) AND (N MOD 10)

x **mod 10**	**Results of** $(x^2 - n)$ **mod 10 when** n **mod 10 equal**			
	1	**3**	**7**	**9**
0	9	7	3	1
1	0	8	4	2
2	3	1	7	5
3	8	6	2	0
4	5	3	9	7
5	4	2	8	6
6	5	3	9	7
7	8	6	2	0
8	3	1	7	5
9	0	8	4	2

Based on the two values, (x mod 10) and (n mod m), we can determine when there is no need to compute x^2. Table II displays the different cases for the relation between (x mod 10) and (n mod 10), see the gray-colored cells. Therefore, the accepted cases to apply the idea of ignoring the squaring operation are as follows [25]: (1) n mod 10=1 and x mod 10=2, 3,7 or 8. (2) n mod 10=3 and x mod 10=0, 1, 4, 5, 6 or 9. (3) n mod 10=7 and x mod 10=0, 2, 3, 5, 7 or 8. (4) n mod 10=9 and x mod 10=1, 4, 6 or 9.

The complete steps of this technique are shown in Algorithm FM1010 [24]. Initially, we determine the initial value of x and (n mod 10). Then the algorithm choose one of the four cases, as described previously. The number of iterations of the FM1010 algorithm is similar to that of the FF algorithm, but some of these iterations ignore the squaring and perfect squaring operations.

Algorithm FM1010 (Fermat Sieving by two Modulus 10)

Input: *n* is a positive odd number.

Output: p and q are two primes such that $n = p\,q$.

```
Begin
 1.  $x = \lfloor\sqrt{n}\rfloor$
 2.  found=false
 3.  $x = x + 1$
 4.  $r_1 = n$ mod 10
 5.  While (Not found) do
 6.      $r_2 = x$ mod 10
 7.      If ($r_1 = 1$) then
 8.          If Not ($r_2 = 2$ or 3 or 7 or 8) then
 9.              $y = x^2 - n$
10.              If ($y$ is a perfect square) then
11.                  found=True
12.              Else
13.                  $x = x + 1$
14.              End if
15.          Else
16.              $x = x + 1$
17.          End if
18.      Else
19.        If ($r_1 = 3$) then
20.          If Not ($r_2 = 0$ or 1 or 4 or 5 or 6 or 9) then
21.                      Similar to lines 9-17
22.        Else
23.          If ($r_1 = 7$) then
24.            If Not ($r_2 = 0$ or 2 or 3 or 5 or 7 or 8) then
25.                      Similar to lines 9-17
26.          Else
27.            If ($r_1 = 9$) then
28.              If Not ($r_2 = 1$ or 4 or 6 or 9) then
29.                      Similar to lines 9-17
30.          End if
31.  End while
32.  $p = x + \sqrt{y}$
33.  $q = x - \sqrt{y}$
End.
```

Similarly, we construct the accepted cases for applying the idea of sieving for the relation between (x mod 10) and (n mod 20) as follows [25]: (1) n mod 20=1 and x mod 10=1, 5 or 9. (2) n mod 20=3 and x mod 10=2 or 8. (3) n mod 20=7 and x mod 10=4 or 6. (4) n mod 20=9 and x mod 10=3, 5 or 7. (5) n mod 20=11 and x mod 10=0, 4 or 6. (6) n mod 20=13 and x mod 10=3 or 7. (7) n mod 20=17 and x mod 10=1 or 9. (8) n mod 20=19 and x mod 10=0, 2 or 8.

Using these cases, we constructed another sieving method named FM1020 algorithm. The pseudocode of the FM1020 algorithm is similar to that of FM1010, except that the While-loop contains eight If-statement, where each If-statement represents a case from the eight cases of the relation between (x mod 10) and (n mod 20) [25].

Remark: Another modified algorithm was proposed by Somsuk, and Tientanopajai [27] and based on analyzing the last m digits of the modulus, where $m \geq 2$. The main drawback of the modified algorithm is the number of different subroutines used is very large when *m* is large.

C. The Fermat without Perfect Squaring Group

The algorithms in this group do not use the perfect squaring operation during the search for the two prime factors. Two main techniques have been proposed for use in integer factorization algorithms that do not employ perfect squaring.

The first algorithm is based on finding two integers *x* and *y* such that the difference between two square numbers is $4n$, i.e., $4n = x^2 - y^2$. This formula can be rewritten as follows [27, 28]:

$$r = x^2 - (y^2 + 4n)$$

The algorithm starts the search with $x = 2\lceil\sqrt{n}\rceil$ and $y = 0$, and then computes the value of *r*. Based on the value of *r*, the algorithm executes one of the following cases [27, 28]:

1) Case 1: $r = 0$. This means that 4*n* is equal to the difference between two squares and the search process is terminated. Therefore, the two prime factors are $p = (x + y)/2$ and $q = (x - y)/2$.

2) Case 2: $r < 0$. This means that the value of the term x^2 is less than the term $(y^2 + 4n)$. Therefore, the algorithm increases the value of *x* by 2 because the value of the term $(p + q)$ is an even number. Also, the algorithm increases the value of *r* by $(4x + 4)$ because this value represents the difference between the squares of the next and the current value of *x*.

3) Case 3: $r > 0$. This means that the value of the term x^2 is greater than the term $(y^2 + 4)$. Therefore, the algorithm increases the value of *y* by 2 because the value of the term $(p - q)$ is an even number. Also, the algorithm decreases the value of *r* by $(4y + 4)$ because this value represents the difference between the squares of the next and current value of *y*.

The complete steps of the algorithm are shown in Algorithm FnPS. All the operations of the algorithm are simple and the running time of the algorithm is based only on the number of iterations of While-loop that is dependent on two conditions: $r < 0$ and $r > 0$. The number of iterations for the first inner while-loop that is based on the condition $(r < 0)$ is $\frac{p+q}{2} - \left(\lfloor\sqrt{n}\rfloor + 1\right)$, while the number of iterations for the second inner while-loop that is based on the condition $(r > 0)$ is $\frac{p-q}{2}$, because the start value of y is 0. Therefore, the total number of iterations of the FnPS algorithm is $p - \left(\lfloor\sqrt{n}\rfloor + 1\right)$.

Algorithm FnPS (Fermat with no Perfect Squares)

Input: *n* is a positive odd number.

Output: p and q are two prime numbers such that $n = p\,q$.

```
Begin
 1.  $x = 2\lceil\sqrt{n}\rceil$
 2.  $y = 0$
 3.  While $(r \neq 0)$ do
 4.      While $(r < 0)$ do
 5.          $r = r + (4x + 4)$
 6.          $x = x + 2$
 7.      End while
 8.      While $(r > 0)$ do
 9.          $r = r - (4y + 4)$
10.          $y = y + 2$
11.      End while
12.  End while
13.  $p = (x + y)/2$
14.  $q = (x - y)/2$
End.
```

The second algorithm in this group is based on removing the perfect squaring operation by using modular multiplication. The method of modification is based on Euler's theorem which is given by the following formula [20]:

$$a^{\Phi(n)} \equiv 1 \bmod n$$

where n is a positive integer number, *a* is an integer such that *a* is relatively prime with *n*, and $\Phi(n)$ is the Euler's totient function that is equal to $(p-1) \times (q-1) = n - (p+q) + 1$.

The complete steps of the algorithm are shown in Algorithm FE [20]. The algorithm starts the computation by initializing $x = 2\lceil\sqrt{n}\rceil$ and selecting a positive integer *c* which is relatively prime with *n*. Based on the value of *c*, the algorithm computes the inverse, *a*, and the square, *s*, of *c* in modulo *n*. Then the algorithm applies Euler's theorem, where the value of $\Phi(n) = n - x + 1$. If the result, *t*, equals 1 then the values of *x* and *y* are determined and the solution is found. Otherwise, the algorithm sets a flag with false. In which case the algorithm repeats the following steps until Euler's theorem is verified, i.e., when *t* = 1. The first step in the repetition is to compute a new value *t* by multiplying it with *s* and taking a module if it is required. The second step is to update the value of *x* by increasing it by a value of 2. Finally, the value of the two prime factors are $x + y$ and $x - y$. The total number of iterations of this algorithm is similar to that of the FF algorithm, but the perfect squaring operation and only one square root operation is required.

Algorithm FE (Fermat-Euler)

Input: *n* is a positive odd number.

Output: p and q are two prime such that $n = p\,q$.

```
Begin
 1.  $x = 2\lceil\sqrt{n}\rceil$
 2.  Choose a positive integer $c$ s.t. gcd$(c, n) = 1$, say $c$=2
 3.  $a = c^{-1} \bmod n$
 4.  $s = c^{2} \bmod n$
 5.  $t = a^{n-x+1} \bmod n$
 6.  If  $(t$ ==1) then
 7.      $x = x/2$
 8.      $y = \sqrt{x^2 - n}$
 9.      found=True
10.  Else
11.      found=False
12.  End if
13.  While (found == False) do
14.      $t = t\,s$
15.      If $(t > n)$ then
16.          $t = t \bmod n$
17.      End if
18.      $x = x + 2$
19.      If $(t == 1)$ then
20.          $x = x/2$
21.          $y = \sqrt{x^2 - n}$
22.          found=True
23.      End if
24.  End while
25.  $p = x + y$
26.  $q = x - y$
End.
```

Remarks:

1) The original statements of Algorithm FnPS in lines 4 and 8 are if $(r < 0)$ and if $(r > 0)$, respectively, see [27]. To optimize the original algorithm, we replaced the two If-statements with two While-loops as in Algorithm FnPS. The performance of Algorithm FnPS with two inner While-loops is better than that with two If-statements, see Section IV.

2) Algorithm EF is slightly different than the algorithm in [20]. In [20], the Boolean condition *found* was not used. Instead, the statement at line 11 is y = 0.1, i.e., any initial value of y is considered to be a float number, and the statement at line 13 is "y is not an integer". The performance achieved by this modification is slightly better than that of the FE algorithm, but not significantly so.

IV. Experimental Evaluations

In this section, we discuss our experimental study of the nine different integer factorization algorithms that utilize Fermat's strategy. The study was conducted according to the methodology described in Section II and involved the evaluation of the performance of these algorithms over different numbers of bits and different differences between the two factors. The first subsection presents the specification of the hardware and software used in the implementation, and the different values of the factors b and Δ that have an effect on the input data. The second subsection presents the results of applying the ideas mentioned in the remarks in Section III. The third subsection presents the measurement and analysis of the running times of the different integer factorization algorithms.

A. Experimental Specification

The experimental study of different factorization algorithms required the use of a high configuration of hardware. For this purpose, we used a Microsoft Azure cloud system. The system is able to run 32 threads in parallel with a processing speed of 2.6 GHz. The reasons for selecting this hardware platform are as follows: (1) The execution times of the factorization algorithms increase rapidly with increases in b and Δ; (2) a large number of instances are used to measure the running time for each algorithm for fixed values of b and Δ; and (3) experimentally, each algorithm is run sequentially, but we use different threads to execute different instances. However, in order to unload the system we used only a maximum of 16 threads.

As regards the implementation, we ran all the algorithms under Windows 2019 and used Microsoft Visual Studio 2019 to implement the algorithms using C++ language.

The specification of the data used in the evaluation of the different factorizing algorithms was as follows:

1) The values of b were 100, 200, 300, 400, and 500.

2) The value of Δ started with (b/4) and then we increased it incrementally by 5 bits until (b/4)+20. The reason that $\Delta = (b/4) + 20$ was set as the last value is that the running time for all the algorithms is very high and the time increases with an increase in b.

3) The value of t, i.e., the number of instances for the fixed values of b and Δ, was set as 100, except when $\Delta = (b/4) + 20$ where we considered t = 25, because the running time for each instance is greater than 1.5 hours for the best algorithm.

B. Experimental Comments

In this subsection, we illustrate the results that were achieved by applying the ideas mentioned in some of the remarks in Section III to optimize some of the statements in the previous algorithms.

However, first, the results in respect of the two versions of the FnPS algorithm, the If-statement and the While-loop, are presented and discussed. Table III shows the results of running two versions of the FnPS algorithm, on b = 100 and Δ =25, 30, 35, and 40. It is clear that the results indicate that the performance of the FnPS algorithm when using the While-loop is better than that when using the If-statement for all values of Δ . On average, the percentage of improvement in the performance of the FnPS algorithm with While-loop was 16.5% compared to its performance with the If-statement.

However, in general, the performance of the FnPS algorithm, even with While-loop, is very weak compared to that of the FF algorithm for two reasons. The first reason, from the theoretical point of view, is that the search space of the FnPS algorithm is very large compared to the search space of the FF algorithm. The second reason, from the practical perspective, is that the running time of the FF algorithm on $\Delta = 25$ and 30 is zero (see subsection IV.C), whereas the running time of the FnPS with the best case is 0.5 seconds. Additionally, the running time of the FF algorithm on $\Delta = \Delta_0 + 15 = 40$ is less than 1 minute, whereas the running time of the FnPS algorithm is greater than 5 hours. Finally, the results for the FnPS algorithm showed that increases in the value of b led to an increase in the search space, so we excluded this algorithm from the full comparison of all the integer factorization algorithms.

TABLE III. PERFORMANCE OF TWO VERSIONS OF FNPS WITH N = 100

Methods	Δ			
	25	**30**	**35**	**40**
FnPS with If-statement	0.6342 s	20.99 s	11.56 m	6.21 h
FnPS with While-loop	0.5146 s	17.57 s	9.73 m	5.25 h
Improvement %	18.86%	16.29%	15.83%	15.46%

Second, the results in respect of the FF1 algorithm are as follows. First, the results of running FF1 algorithm, on b = 100 and Δ =25 and 30, 35 are 0.001 and 0.002 seconds, respectively. Second, the results of running FF1 algorithm, on b = 100 and Δ =35 is greater than one hour in many cases without find the solution. As we see in the Subsection C, the running time for FF algorithm is faster than FF1 algorithm when $b \geq 100$, so we excluded this algorithm from the full comparison of all the integer factorization algorithms.

C. Results of the Comparison

Based on the results reported in Subsection IV.B, we compared seven of the nine algorithms, i.e., FF, FM10, FM20, FM1010, FM1020, FOE, and FE. The other two algorithms, FF1 and FnPS, were excluded from the final comparison because their performance was very poor.

The results of implementing the seven integer factorization algorithms using the data sets and platform described above are shown in Fig. 1, 2, 3, 4, and 5 for n = 100, 200, 300, 400, and 500, respectively. From the results shown in the figures, we can make several observations.

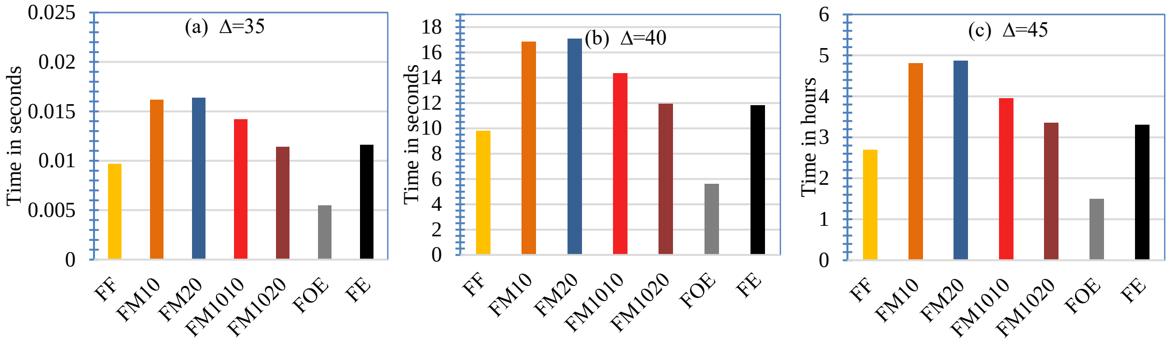


Fig. 1. Running Time for Fermat Algorithms when *n*=100.

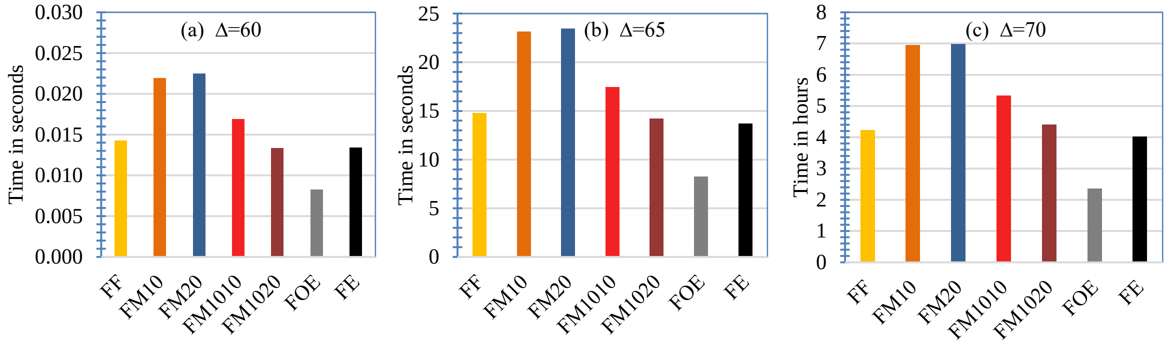


Fig. 2. Running Time for Fermat Algorithms when *n*=200.

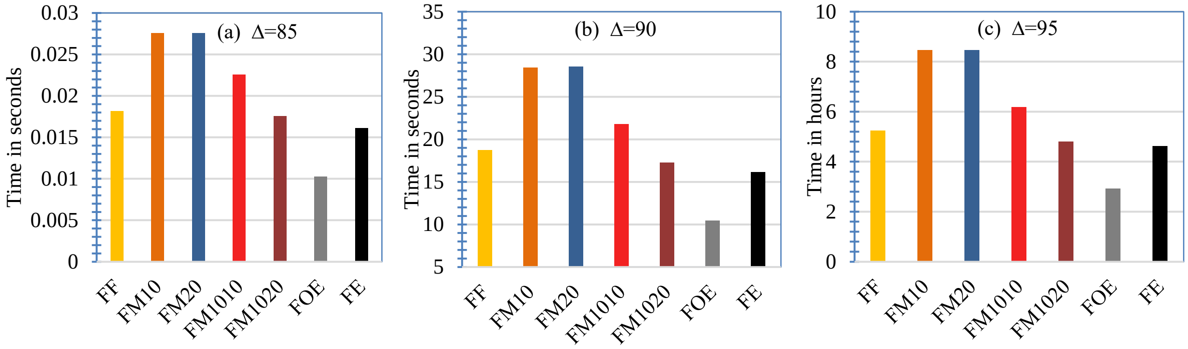


Fig. 3. Running Time for Fermat Algorithms when *n*=300.

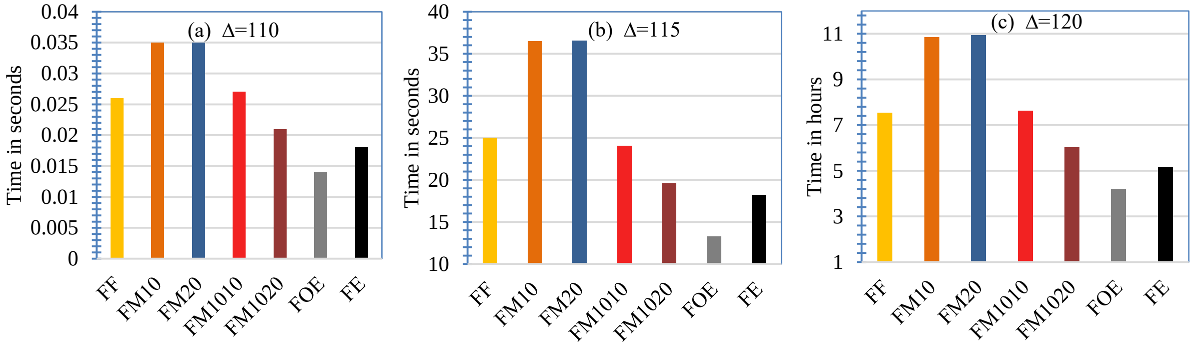


Fig. 4. Running Time for Fermat Algorithms when *n*=400.

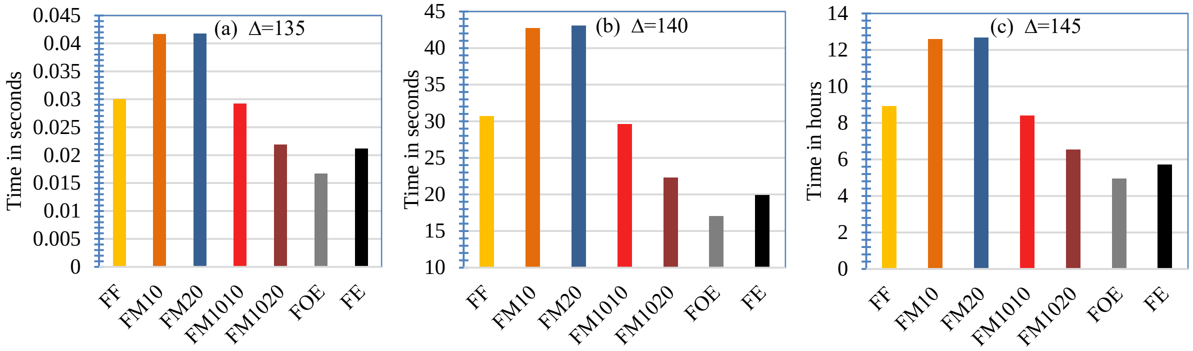


Fig. 5. Running Time for Fermat Algorithms when n=500.

First, the running time of all the algorithms is 0 seconds for all values of b and when $\Delta = b/4$, and $(b/4) + 5$, except for the FE algorithm which has a very short runtime of 0.001 seconds when b = 400 and 500. For this reason, there are no data in the figures for $\Delta = b/4$, and $(b/4) + 5$. This means that the running time for all seven of the Fermat-based algorithms is instantaneous in the case of $\Delta = b/4$, and $(b/4) + 5$.

Second, for a fixed value of b with increasing values of Δ, the running time for all seven algorithms rapidly increases. We can estimate this factor by calculating the percentage of the increase in the running time between two consecutive Δ, Δ_1 and Δ_2, and a fixed value of b, where $\Delta_2 = \Delta_1 + 5$. By way of an example, Table IV shows this ratio for the FE algorithm when b = 300. The running times for the FE algorithm when b = 300 and Δ =85, 90, and 95 are 0.0161 seconds, 16.18 seconds, and 4.62 hours, respectively. The percentage increase in the running time when Δ_2=90 is approximately equal to 1000 times the running time when Δ_1 = 85, and the percentage of increase in the running time when Δ_2 = 95 is approximately equal to 1000 times the running time when Δ_1 = 90. This phenomenon is true for all values of b and for all seven of the integer factorization algorithms studied.

Third, the difference between the FM10 and FM20 algorithms is not significant and the two algorithms are near to equal for all values of b and Δ. On the other hand, the difference between the FM1010 and FM1020 algorithms is significant in respect of the average case. The FM1020 algorithm has better performance than the FM1010 algorithm by a percentage of 15.17%, 17.31%, 22.33%, 21.10%, and 22.17% for n = 100, 200, 300, 400, and 500, respectively. This means that the average percentage of improvement achieved by FM1020 is 19.62% when compared to FM1010.

Fourth, the performance of the FM10 and FM20 algorithms is weak compared to the other five algorithms for every b and Δ. The original Fermat algorithm, FF, has better performance than the FM10 and FM20 algorithms by almost 36% in respect of the average case.

Fifth, Table V shows the percentage of improvement achieved by each algorithm A compared to the FF algorithm for all values of b, where algorithm A is one of the following algorithms: FM10, FM20, FM1010, FM1020, FOE, and FE. Two types of value are presented in the table. The first value is a positive value, say ∝, which means that algorithm A has ∝ improvement compared to the FF algorithm, and the second value is a negative value, say -∝, which means that the FF algorithm has ∝ improvement compared to algorithm A.

Sixth, from Table V, it is clear that the percentage of disimprovement of algorithm A decreases with an increase in the number of bits. For example, the FM1010 algorithm has a disimprovement of 32.06%, 20.61%, 15.21%, and 1.04% for n = 100, 200, 300, and 400, respectively. Also, the percentage of improvement of algorithm A increases with an increase in the number of bits, except for the FOE algorithm. For example, the percentage of improvement achieved by the FE algorithm when compared to the FF algorithm is 4.89%, 11.83%, 31.78%, and 35.97% for n = 200, 300, 400, and 500, respectively.

Seventh, the FOE algorithm has two properties compared to the FF algorithm. The first is that the FOE algorithm has better performance compared to the FF algorithm for all values of b. The second is the percentage of improvement of the FOE algorithm compared to the FF algorithm is almost fixed and equal to 44% for all values of b. The reason for this property is that the number of iterations of the FOE algorithm is half that of the FF algorithm.

TABLE IV. RATIO OF RUNNING TIME BETWEEN TWO CONSECUTIVE Δ WHEN B = 300 FOR EACH ALGORITHM

Δ_2/Δ_1	**Fermat Algorithms**						
	FF	**FM10**	**FM20**	**FM1010**	**FM1020**	**FOE**	**FE**
90/85	1027	1029	1034	964	979	1016	1002
95/90	1009	1072	1068	1019	1001	1004	1031

TABLE V. PERCENTAGE OF IMPROVEMENT ACHIEVED BY EACH ALGORITHM COMPARED TO FF ALGORITHM

b	**Fermat Factorization Algorithms**					
	FM10	**FM20**	**FM1010**	**FM1020**	**FOE**	**FE**
100	- 44.12	- 44.80	- 32.06	- 19.91	43.99	- 18.70
200	- 39.12	- 39.42	- 20.61	- 3.99	44.38	4.89
300	- 38.06	- 38.13	- 15.21	8.40	44.27	11.83
400	- 30.42	- 30.92	- 1.04	20.27	44.24	31.78
500	- 29.27	- 29.61	5.87	26.73	44.38	35.97

Eighth, Fig. 6 displays the behavior of all the algorithms for the average case for all different values of b. The same behavior occurs if we fix the value of Δ and change the value of b. It is clear that the best integer factorization algorithm based on Fermat's strategy for all data sets is the algorithm that is based on sieving odd and even numbers, i.e., the FOE algorithm. The second and third best algorithms are FE and FM1020, respectively. Note that the two curves for the FM10 and FM20 algorithms coincide.

Ninth, Fig. 7 displays the running times for $t = 100$ instances for each algorithm A when $b = 500$ and $\Delta = 140$, where algorithm A is one of the following algorithms: FM1010, FM1020, FOE, and FE. From the figure, we can observe the following: (1) The running time of each algorithm A is varied and based on the value of n, except for the FE algorithm whose behavior is almost fixed; and (2) the running time of the FM1010 algorithm is slightly faster than that of the FF algorithm on average, as in Fig. 5(b), but in many instances the running time of the FM1010 algorithm is longer than the FF algorithm.

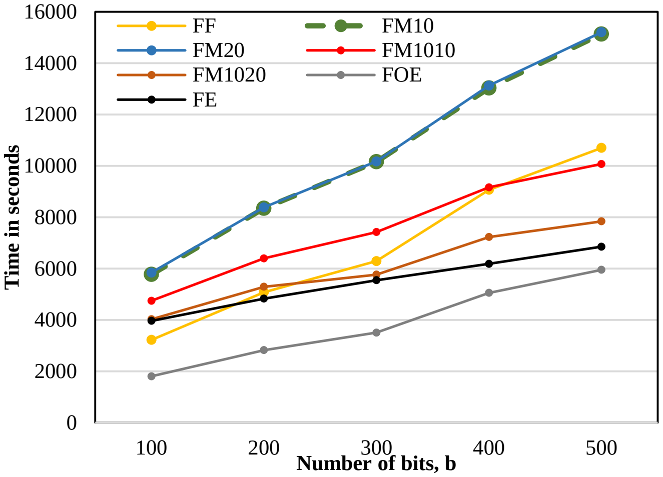


Fig. 6. Average Running Time of Fermat algorithms over Different *n*.

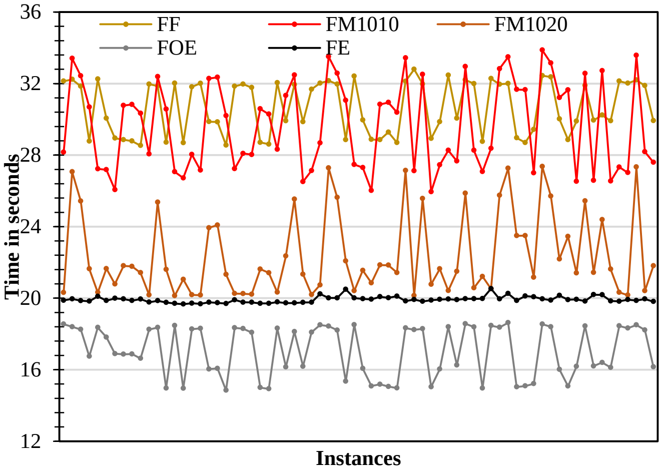


Fig. 7. Behavior of FF, FM1010, FM1020, FOE and FE algorithms when b = 500 and Δ= 140.

V. Conclusion and Open Problems

In this work, we addressed the integer factorization problem that involves finding the two prime factors for an odd composite number. The inherent challenge of this problem is that no polynomial time has yet been found. The problem is used in cryptography to break the RSA cryptosystem. We classified the different Fermat factorization algorithms into three groups according to the integer factorization methods that were used. The algorithms were studied experimentally to measure their performance according to two factors: (1) the size of the odd composite factor, b bits, and (2) the difference between two factors, Δ bits. The experimental study was conducted on data sets consisting of b =100, 200, 300, 400, and 500, and $\Delta = b/4$, $(b/4) + 5$, $(b/4) + 10$, $(b/4) + 15$, and $(b/4) + 20$. The results of the experimental study showed that the algorithm based on sieving using the odd and even property performed the fastest factorization when applied to data sets of different sizes of odd composite numbers and different differences between the two factors with a percentage improvement of 44% compared to the original Fermat factorization algorithm. Also, the algorithm based on Euler's theorem exhibited a good level of performance compared with the Fermat factorization algorithm.

There are open questions remain related to this study. (1) What is the behavior of FM1020, FOE and FE algorithms when $b > 500$ bits? (2) How to use high-performance computing such as graphics processing unit (GPU) to speedup the running time for the best modified Fermat algorithm. (3) What is the effect of using n mod m on FM1020 algorithm, when m>20.

Acknowledgment

This work has been funded by Scientific Research Deanship at the University of Ha'il – Saudi Arabia through project number RG-191309.

References

[1] K. Balasubramanian, and M. Rajakani, "Algorithmic strategies for solving complex problems in cryptography," IGI global, 2017.

[2] S. Yan, "Primality testing and integer factorization in public-key cryptography," Springer, 2009.

[3] M. Daniel, "A survey of fast exponentiation methods," J. Algorithm vol. 27, no. 1, pp. 129–146, 1998.

[4] K. Fathy, H. Bahig, and A. Ragab, "A fast parallel modular exponentiation algorithm," Arabian Journal for Science and Engineering, vol. 43, pp. 903–911, 2018.

[5] R. Crandall, and C. Pomerance, "Prime numbers: a computational perspective," 2nd Ed., Springer, 2005.

[6] O. Akchiche, and O Khadir, "Factoring RSA moduli with primes sharing bits in the middle," Applicable Algebra in Engineering, Communication and Computing, vol. 29, pp. 245–259, 2018.

[7] H. Bahig, D. Nassr, and A. Bhery, "Factoring RSA modulus with primes not necessarily sharing least significant bits," Applied Mathematics and Information Sciences, vol. 11, no. 1, pp. 243-249, 2017.

[8] H. Bahig, D. Nassr, A.Bhery and A. Nitaj, "A unified method for private exponent attacks on RSA using lattices," International Journal of Foundations of Computer Science, vol. 31, no. 2, pp. 207-231, 2020.

[9] R. Steinfeld, and Y. Zheng, "On the security of RSA with primes sharing least-significant bits," Applicable Algebra in Engineering, Communication and Computing, vol. 15, pp. 179–200, 2004.

[10] A. Lenstra, "Integer factoring," Designs, Codes and Cryptography, vol. 19, pp. 101–128, 2000.

[11] J Jormakka, “On finding Fermat’s pairs,” Journal of Discrete Mathematical Sciences & Cryptography, vol. 10, no. 3, pp. 401-413, 2006.

[12] B. de Weger, “Cryptanalysis of RSA with small prime difference,” Applicable Algebra in Engineering, Communication and Computing, vol. 13, no. 1, pp. 17–28, 2002.

[13] H. Bahig, H. Bahig, and Y Kotb, “Fermat factorization using a multi-core system,” International Journal of Advanced Computer Science and Applications, vol. 11, no. 4, pp. 323-330, 2020.

[14] R. Erra, and C. Grenier, “The Fermat factorization method revisited,” IACR Cryptology ePrint Archive, 318 2009.

[15] W. Hart, “A one line factorization algorithm,” Journal of the Australian Mathematical Society, vol. 94, pp. 61-69, 2012.

[16] M. Hittmeir, “Deterministic factorization of sums and differences of powers,” Mathematics of Computation, vol. 86, no. 308, pp. 2947–2954, 2017.

[17] J. Mckee, “Speeding Fermat’s factorization method,” Mathematics of Computation, vol. 68, no. 228, pp. 1729–1737, 1999.

[18] B. Randall, “Fingers find Fermat’s factorization most probable,” The Mathematical Gazette, vol. 99, no. 544, pp. 452-458, 2014.

[19] P. Shiu, “Fermat's method of factorization,” The Mathematical Gazette, vol. 99, no. 544, pp. 97-103, 2015.

[20] K. Somsuk, “The new integer factorization algorithm based on Fermat’s factorization algorithm and euler’s theorem,” International Journal of Electrical and Computer Engineering, vol. 10, no. 2, pp. 1469-1476, 2020.

[21] K. Somsuk, and S. Kasemvilas, “MVFactor: A method to decrease processing time for factorization algorithm,” Proceedings of 17th International Computer Science and Engineering Conference, Thailand, 2013, pp. 339-342.

[22] K. Somsuk and S. Kasemvilas, “MFFV2 and MNQSV2: improved factorization algorithms,” Proceeding of 4th International Conference on Information Science and Applications, 2013, pp. 327 – 329.

[23] K. Somsuk and S. Kasemvilas, “Possible prime modified Fermat factorization: new improved integer factorization to decrease computation time for breaking RSA,” Proceedings of the 10th International Conference on Computing and Information Technology. Advances in Intelligent Systems and Computing, vol. 265, pp. 325-334, 2014.

[24] K. Somsuk and S. Kasemvilas, “MFFV3: An Improved Integer Factorization Algorithm to Increase Computation Speed”, 5th International Engineering Conference 2014, pp. 1432 – 1436, 2014.

[25] K.Somsuk, “A new modified integer factorization algorithm using integer mod 20's technique,” Proceedings of the 18 International Computer Science and Engineering Conference, Thailand, 2014, pp. 312-316.

[26] K. Somsuk and K. Tientanopajai, “Improving Fermat factorization algorithm by dividing modulus into three forms,” KKU Engineering Journal, vol. 43, no. S2, pp. 350-353, 2016.

[27] K. Somsuk, and K. Tientanopajai, “An improvement of Fermat's factorization by considering the last m digits of modulus to decrease computation time,” International Journal of Network Security, vol. 19, pp. 99-111, 2017.

[28] M. Wu, R.Tso, and H. Sun, “On the improvement of Fermat factorization using a continued fraction technique,” Future Generation Computer Systems, vol. 30, no. 1, pp. 162-168, 2014.

[29] G.Xiang, “Fermat’s method of factorization,” Applied Probability Trust, vol. 36, no. 2, pp. 34-35, 2004.

[30] R. Sakellariou, “Parallel algorithms for integer factorization,” Advances on Computer Mathematics and Its Applications, pp. 288-295, 1993.

[31] G. Hiary, “A deterministic algorithm for integer factorization,” Mathematics of Computation, vol. 85, pp. 2065-2069, 2016.

[32] E. Costa, and D. Harvey, “Faster deterministic integer factorization,” Mathematics of Computation, vol. 83, pp. 339-345, 2014.

[33] R. Sakellariou, “Parallel algorithms for integer factorization,” Advances on Computer Mathematics and Its Applications, pp. 288-295, 1993.

[34] GMP library, The GNU multiple precision arithmetic library. https:// gmplib.org/, 2020.

[35] H. M. Bahig, A. Alghadhban, M. A. Mahdi, K. A. Alutaibi, H. M. Bahig, “Speeding up the multiplication algorithm for large integers”, Engineering, Technology & Applied Science Research, vol 10, no. 6, pp 6533-6541, 2020.

[36] H. Bahig, H. Bahig, and K. Fathy, “Fast and scalable algorithm for product large data on multicore system,” Concurrency and Computation: Practice and Experience, https://doi.org/10.1002/cpe.5259, online published 2019.

Agent-based Model for Simulating Urban System

Fatimazahra BARRAMOU[1], Malika ADDOU[2]
LaGeS Laboratory, Hassania School of Public Works
Casablanca, Morocco

***Abstract*—In this paper, we address the issue related to the modelling and simulation of urban systems. We propose a new approach for simulating urban system based on multi-agent paradigm. Our proposed model is based on the use of a coupling between cellular agents and vector agents. These agents made it possible to take into account the spatial dimension of urban systems as well as the modelling of all the rules that govern them. To allow reusability of our model, we apply the VOYELLE approach by defining an environment model, an organization model, an agent model and an interaction model. We test our proposed model with a case study on Casablanca city. We discuss the problem of urbanization of Casablanca by following an approach that reduces the problem into two sub-problems similar but are treated differently: first, predict the city's need of housing (individual housing zone, multifamily housing zone, ...) and then anticipate the city's need of public services and ensure better spatial distribution of these equipment to best serve the people needs. Then we did experimentation with two simulation scenarios by changing in each scenario the hypotheses concerning urban planning especially in terms of demographic growth rate and residential sprawl.**

Keywords*—*Multi agent systems; simulation; modelling; urban system; cellular agent; vector agent

I. Introduction

In the literature, there are several simulation techniques of complex systems: micro simulation [1], cellular automata [2] [3] and the object-oriented [4] simulation. However, when the system reaches a certain level of complexity, these techniques are insufficient, first to model the system dynamics and the interactions between the elements, and secondly, these techniques do not allow the system representations at different levels (micro / macro) and does not explain the emergence of space-time structures.

To overcome these problems, a new trend has emerged in recent years: the agent-oriented simulation [5] [6] [7] that took considerable development and it is now used in a growing number of sectors, where it is gradually replacing other simulation techniques. Indeed, the multi-agent systems can simulate systems with spatial component by representing the spatial dimension with the agent's property "situated" in an environment and so they can represent the space-time environment where agents are evolved. They also offer the ability to model the endogenous dynamics of the system and the interactions between the different agents. Some studies [8] used the concept of cellular agent to describe the geographical area as a cell. More recent studies use the vector agent [9] that describes the space along a well-defined geographical form (point, line or polygon).

In this paper, we use a new approach based on a coupling between both the cellular agents and vectors agents to describe and simulate the spatial systems. We focus on the urban system that can be considered as a spatial system in which several elements involved in its organization.

II. Background of the Study

A. Multi-agent Paradigm

Agent is defined by Ferber [10] as autonomous entity, real or abstract, which is able to act on itself and on its environment. In a multi-agent universe, this entity can communicate with other agents, and whose behavior is the consequence of its observations, knowledge and interactions with other agents. Therefore, messages between agents differ from messages between objects because:

- Agent is autonomous and decides on his own process whether or not to perform a required action.
- Agents have their own goals and act proactively to achieve theme.
- Agents are capable of social behavior: engaging in complex interactions.

A multi agent system is composed of the following elements [10]:

- An environment E, it's a space generally having a metric.
- A set of objects O. These objects are located. For any object, it is possible to associate a position in E. These objects are passive; they can be perceived, created, destroyed and modified by the agents.
- A set A of agents which are particular objects, which represent the active entities of the system.
- A set of relations R which unite objects (and therefore agents) between them.
- A set of operations Op allowing agents of A to perceive, produce, consume, transform and manipulate objects of O.
- Operators responsible for representing the application of these operations and the reaction of the world to this attempt at modification, which will be called the laws of the universe.

To design a multi-agent system, it is not enough to place several agents in the same environment, it is also necessary that these agents interact. The analysis and development of a multi-agent system require taking into account multiple dimensions. In the Vowel AEIO approach [11], four dimensions are identified: Agent, Environment, Interactions, and Organization.

B. Simulation Theory

Simulation is a very active branch of computer science that involves analyzing the properties of theoretical models of the surrounding world. According to Ferber, numerical simulation is a representation of reality based on the underlying theoretical model [12]. The goals of the simulation are:

- Better understanding of the system;
- System performance measurement;
- Sizing / optimization of production systems (before / after completion);
- Identification of critical factors;
- Response to the questions "What happens ... if ...?";
- "Real time" decision support tool (during operation);
- Staff training tool (flight simulators, choice simulators, etc.).

Simulation techniques have been applied in several fields, especially in the urban field. The term urban simulation is used to denote the attempt to reproduce urban dynamics [13]. These are often simulation approaches based on cellular automata that are proposed in the literature [14], but we also find some cases which deal with the problem with agent-based models [15]. Modeling and simulation then concern the various processes relating to urban development.

Geosimulation, is a recent sub-category of urban simulation in which we manipulate individuals (humans) and infrastructure entities at unmodifiable spatial scales such as houses, households and plots [16] [17]. An urban micro-simulation is not necessarily a geosimulation. For example, in Urbansim [18], as well as in the work of Miller's team [19], the simulation of urban systems was done using representations of elementary units at the microscopic scale, except that these units are not spatial objects [20]. The term "micro" is used here to designate data disaggregated on a regular partition of space at a certain scale defined by the needs of the model, but not necessarily that relating to the objects of the analysis.

III. Simulation of Spatial Systems

The modeling and simulation of spatial systems have been developed in the literature using several approaches. These include cellular automata, cellular agent and vector agent.

A. Simulation by Cellular Automata

The concept of Cellular Automata (CA) has been defined as the interconnection in a space of a set of automata with finite number of states. It is a regular grid of cells, each cell can be found in a finite number of discrete states with no discrete time during its evolution. The principle of the cellular automaton is the following:

- To each cell is associated a state, described by one or more variables [21];
- At each discrete time, the same rules are applied simultaneously to all the cells of the grid: the new state of a cell depends on its state at the previous time and state of the cells in its neighbourhood [3];
- The description of the operation of a cell is given by a graph called a state diagram. All cells are built on the same pattern. The state diagram is characteristic of the entire cellular automaton.
- The definition of a cellular automaton is relatively simple but the result of his behaviour can be very complex.

The cellular automata approach has been very successfully used in simulating complex systems. Indeed, the CA allow the modeling of complexity, they are flexible, they are easy to interpret, they have a certain affinity with geographic information systems (GIS) due to their nature "raster" and are relatively simple tools [3]. Despite the benefits of the cellular automata, they are not adapted to simulate the spatial component of complex systems due to their low correspondence with the spatial and urban theory and lack of attention to detail [3]. Indeed, spatial complex systems integrate a wide variety of space objects that interact non-linearly and with a complexity that exceeds the cellular automata. Thus, some researchers have considered it’s necessary to resort to other approaches to model the spatial dynamics and in particular cellular agents [22] defined below.

B. Simulation by Cellular Agent

Ferber in [12] considers Cellular Agents (CAg) as special multi-agent systems where cellular agents are surface elements contiguous and fixed. Cellular agent has a goal to fill with spatial characteristics. It is able to access and manage spatial information to solve the appropriate tasks [8].

Unlike cellular automata, the cellular agents occupy positions in a two-dimensional grid of cells and the distance between them affect their interactions [23]. This is called intelligent, autonomous cells that interact to achieve their own goals [24]. Simulating cellular agent also has the advantage of integrating cellular agents with geographic information system [25].

Several studies have used the cellular agent’s approach [14, 26, 27, 28, 29]. However some of these works have shown some limitations of cellular agents. Indeed, cellular agents are limited by the shape of the cells that do not provide a good model of irregular objects in space and neighbourhood relations. Thus, some authors have opted for a different approach to model geometric forms of space, which is the vector agent.

C. Simulation by Vector Agent

A multi-agent system with spatial component may include a vector modelling space. In this case each spatial agent has no limit in shape and size [30]. We call such agent a vector agent (VAg). Vectors agents allow for more realistic modelling of geographic space as cellular automata or cellular agents. Indeed, CA and CAg represent the geographic space into several cells while VAg may represent a definite spatial entity [31]. In addition VAg may have an infinite number of neighbours without distance limitations [31]. In the case of the cellular automata, transition functions must be applied only to the immediate neighbours (typically 8 cells) and a cell cannot be affected by neighbours being at a greater distance. This simulation technique by VAg can be a good approach to reproduce the classic urban theories, each class of space objects with its own transition rules and neighbourly relations that can be defined in several ways depending on the real world. The VAg offers the ability to change the geometry by simply altering parameters [32]. Finally VAg can represent realistically the complexity of geographical space and simulate the interaction of objects in the space [33].

Few studies to date are based on the vector agent approach. For example, those of the IGN (French National Geographic Institute) are among the most remarkable in the field of automatic cartographic generalization [34, 35, 36, 37]. In [38] Researchers have developed a prototype in which map objects become "cartographic agents" have methods to manage conflicts of mapping example.

IV. Proposed Approach

A. Simulation Objectives

By studying the urban system, we detected two main challenges facing the city:

- The habitat: the city must control urbanization and housing. Indeed, during the urban planning of the city, the residential areas to create must be defined in a clear urbanization policy to install the new population.
- The public services: the objective of urban planning is to make cities well equipped by a set of public services and facilities (schools, hospitals, sports grounds, etc.) to meet the needs of the inhabitants. Also, these public services must be well distributed spatially to ensure equitable access to the entire population.

To simulate the evolution of the urban system, we followed an original approach, which consists in basing the analysis on the examination of concepts associated with the processes of population growth and urban development:

- The demographic growth rate: this parameter defines the evolution of the population within cities. This rate may change from one municipality to another;
- Annual residential sprawl: this is a parameter that makes it possible to calculate and estimate the way in which the built urban space evolves or is created. It is calculated from statistics on the authorized projects in the city;
- Public services: depending on the territory studied, there are standards that specify the number of facilities to be created per number of inhabitants and their areas (for example: create a school for each 8,000 inhabitant with a minimum surface area of 4000m²).

The objective of our simulation model is to present a tool that helps in the urban planning, by simulating the city according to three aspects:

1) *Housing aspect:* Simulate the needs of the city in terms of housing (individual housing, collective housing ...);

2) *Public services aspect:* Simulate the needs of the city in terms of public services (schools, hospitals, sports field, mosque, etc.);

3) *Spatial equality aspect:* Simulate the best spatial distribution of public services to guarantee equitable accessibility to the entire population.

B. Proposed Model

To simulate the evolution of the urban system, we propose an approach based on the coupling between cellular agents and vector agents as shown in Fig. 1.

- The cellular agent will represent the continuity of the geographical space by a two-dimensional grid of contiguous and fixed agent cells.
- The vector agent will represent the geographic objects in space with a well-defined geometric shape in the form of a point, line or area.

In the Fig. 2, we present an agent based model for simulating urban systems. The advantage of the use of Cellular agent and vector agent is to benefit from these two types of agents to represent the urban space. Indeed, the cellular agents will allow representing the spatial continuity of geographical space. Each cellular agent will have information about the land use. While the vector agents will allow to represent the environment's vector objects (as point, line or polygon). For example, we can represent by t agent urban road, public services, schools.

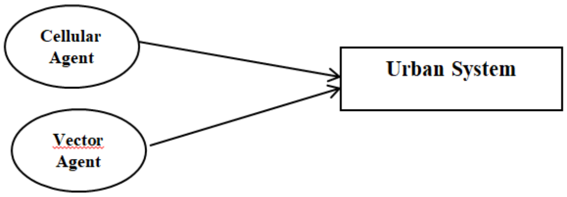


Fig. 1. Coupling Cellular Agent and Vector Agent.

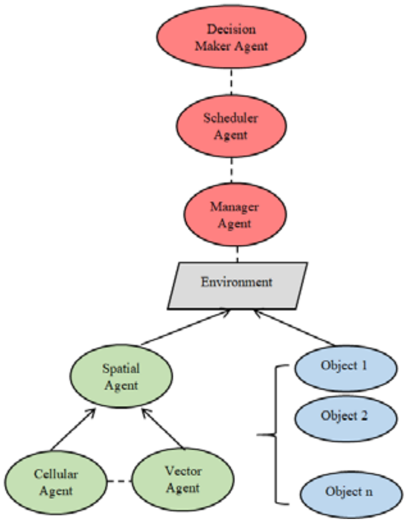


Fig. 2. Agent based Model for simulating Urban Systems.

To enable reuse of our model, we will apply the VOWEL approach [10] to describe the modelling of the four components of multi-agent systems include: the environment, the organization, the agents and the interactions.

C. Modelling of Environment

We consider that the environment is a heterogeneous environment composed by agents and objects. Agents will perceive, communicate and evolve while the objects will simply produce influences in a fixed radius. The environment in which the agents evolve is a continuous Euclidean space. It's a space with a geographic reference system and projection system. Fig. 3 presents the environment modelling.

The environment of our system is the urban system composed with:

- Spatial agents: represent cellular agent and vector agent of the system that communicate and interact to achieve the simulation.
- External agents: represent non-spatial agents that interact to achieve the simulation.
- Objects: represent all spatial object of the urban space.

D. Modelling of Organization

According to Ferber [5], the organization is both the medium and the manner how are the interrelations between the agents. We mean by interrelation delegation of tasks, information transfer, commitments, timings of shares, etc. To study the organization of our system, we will apply the Ferber approach which is based on three components: organizational levels, functional analysis and structural analysis.

1) Organization levels: We modelled the urban system in three levels as shown in Fig. 4:

- Macro level: This level corresponds to the city as a whole consists of communes and districts.
- Meso level: This level corresponds to the communes or districts, which consists of a cell aggregation.
- Micro level: Contains the basic element of the simulation "cell". Each cell has a state (nude, built ...) and potential (urban potential, industrial potential…) and may optionally include a service.

2) Functional analysis: In order to achieve objectives of the organization, all activities that agents are supposed to exercise should be described. This is called Role. To model the roles of an organization, we can either explore the role from the point of view of the agent that acts (agent centred) or from the point of view of the organization (organization centred).

- Agent centred view: In this view the agent is seen as a stable entity to which roles are temporarily attached.
- Organization centred view: In this vision, the role is a stable entity and agents are attached to it.

In the context of our research, we chose the method of agent centred view. Indeed, we will define the set of roles (functions) that each agent plays. We present in Table I the roles of agents in our simulation model.

3) Structural analysis: The goal of structural analysis is to give an order to all the possible interactions between agents by identifying the abstract relationships that link them, and how they evolve over time. It define the relationship between the function of the agent and the function of the organization to which it belongs. We present in Table II the various relationships between the agents of the simulation model:

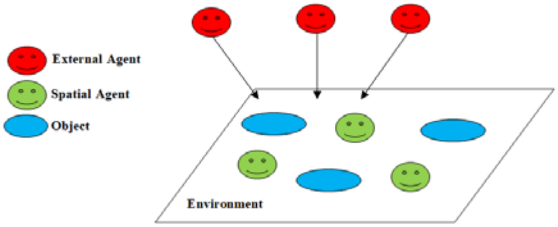


Fig. 3. Modelling of environment.

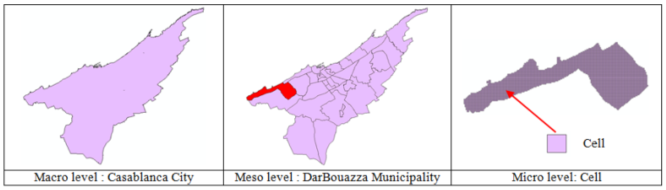

Cell
Macro level : Casablanca City
Meso level : DarBouazza Municipality
Micro level: Cell

Fig. 4. Example of Organization Modelling.

TABLE I. AGENTS ROLES

Agent	Role	Designation
Manager Agent :	Mediator: Performer: Archivist:	It manages the execution requests and distributes them to the competent agents. It performs a set of actions. It memorizes events on the environment.
Scheduler Agent :	Planner: Scheduler :	It determines the actions to execute. It determines the sequence of actions to execute.
Decision Maker Agent :	Performer: Decision-maker:	It runs the simulation. It chooses from a set of possible actions.
Cellular Agent :	Provider : Performer: Decision-maker:	It is rendering services to other agents. It performs a set of actions. It chooses to change its state or not.
Vector Agent :	Performer: Decision-maker:	It performs a set of actions. It chooses the best location.

TABLE II. AGENTS RELATIONSHIP

Relationship	**Agents**
Relationship of acquaintance	All agents
Communication relationship	All agents
Subordination relationship / static	Manager Agent – Cellular Agent
	Manager Agent – Vector Agent
	Decision Maker Agent – Manager Agent
	Scheduler Agent – Manager Agent
Operative relationship	Vector Agent – Cellular Agent
	Scheduler Agent – Manager Agent
Informational relationship	Manager Agent – Vector Agent
	Cellular Agent – Manager Agent
	Vector Agent – Manager Agent
Conflictual relationship	-----
Competitive relationship	Cellular Agent – Vector Agent

E. Modelling Agents

To model the agents of our system, we opted for the AUML approach:

- Manager Agent: It is a reactive agent. This agent is responsible for the initialization of the simulation. It creates and acts on all environmental agents. It interacts with the "scheduler" agent to determine the scheduling of its acts on the agents as shown in Fig. 5.

- Cellular Agent: This is the basic agent of the simulation system. It is a cognitive agent; its decisions are motivated by specific goals that reflect a perception and a logical representation of its environment. It is the spatial entity that constitute the environment, it has: a "state" attribute that contains information about the land use can take the following values (bare land occupied land), a "zonning" attribute which contains the allocation of zoning (residential zoning, industrial zoning,…), potentiel_urban attribute, potentiel_equip attribute, potentiel_environnement attribute, these attributes vary between 0-3 and indicate the cell potential to be intended for urbanization, receiving public service or to be kept in green space, … as shown in Fig. 6.

- Vector Agent: It is a cognitive agent that has knowledge about its environment and reacts to this environment. The simulation technique by cognitive agent is a good approach to find the best location of each vector agent in the urban system. It has goals to accomplish for its own satisfaction (the best location) and a set of behavioral rules that dictate the way it interacts with other agents in the system as shown in Fig. 7.

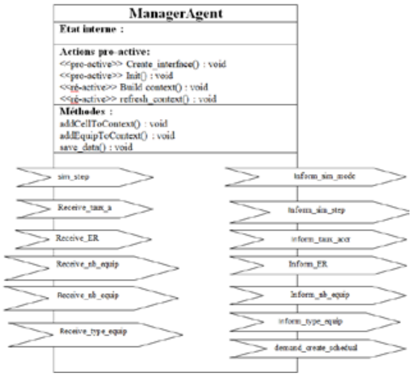


Fig. 5. Manager Agent Conception in AUML.

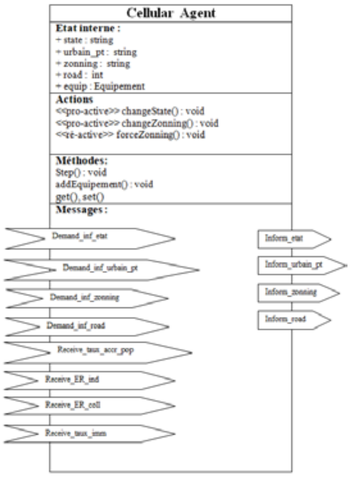


Fig. 6. Cellular Agent Conception in AUML.

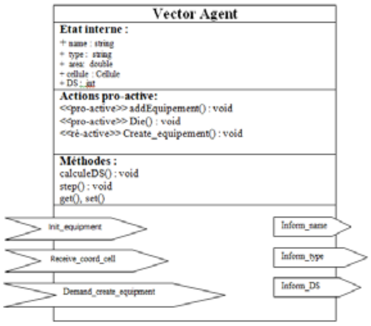


Fig. 7. Vector Agent Conception in AUML.

- Scheduler Agent: It is a reactive agent. Its role is to schedule the actions of Decision Maker Agents in the environment as shown in Fig. 8.

- Decision Maker Agent: It is a cognitive agent. It represents the user of the simulator that can proceed different simulation scenarios by changing the simulation parameters as shown in Fig. 9.

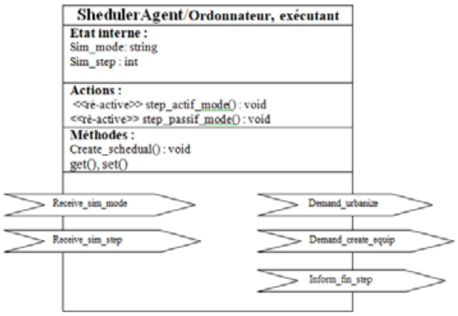


Fig. 8. Schedular Agent Conception in AUML.

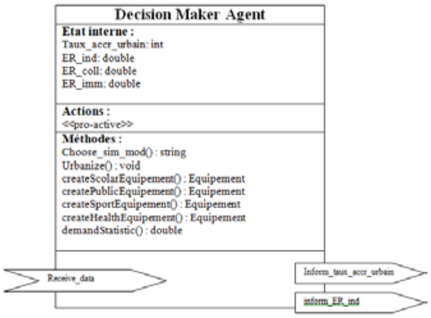

Decision Maker Agent
Etat interne :
Taux_accr_urbain: int
ER_ind: double
ER_coll: double
ER_imm: double
Actions :
<<pro-active>>
Méthodes :
Choose_sim_mod() : string
Urbanize() : void
createScolarEquipement() : Equipement
createPublicEquipement() : Equipement
createSportEquipement() : Equipement
createHealthEquipement() : Equipement
demandStatistic() : double
Receive_data
Inform_taux_accr_urbain
inform_ER_ind

Fig. 9. Decision Maker Agent Conception in AUML.

F. *Proposed Architecture*

To implement our simulation model, we chose the platform Repast Simphony [39] for multi agent system and ArcGIS [40] for GIS data. Fig. 10 presents our proposed architecture:

The proposed architecture consists of the following layers:

- Presentation Layer: It is the part of the visible and interactive application with users. Repast Symphony's plugin uses the Application Framework (SAF), which provides a set of plugins for the configuration of the interface.
- Occupation Layer: It corresponds to the functional part of the application, the application that implements the "logic", which describes the operations that the application operates on data based on user requests made through the presentation layer.
- Data Access Layer: It consists in managing party access to system data. Repast simphony uses Freeze dryers module that provides persistence in xml and text formats (allows you to send and receive the status of the simulation from a secondary storage such as shp file, xml file, ...).
- Data storage: data is stored either in Geodatabase or in shapefile file (.shp).

In Fig. 11, we present the project environment in Eclipse.

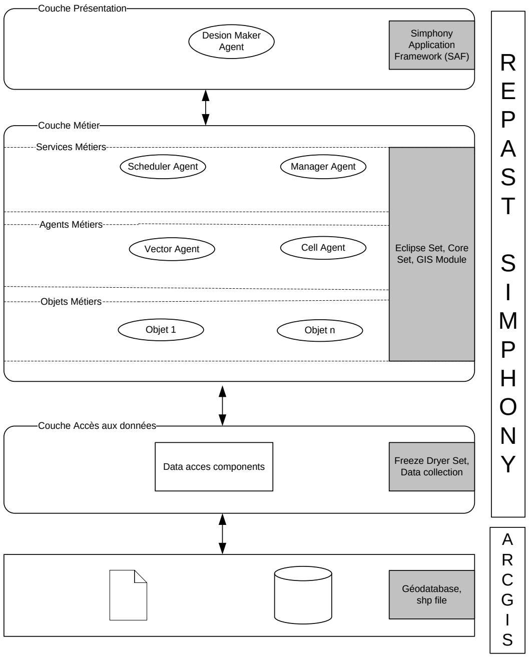
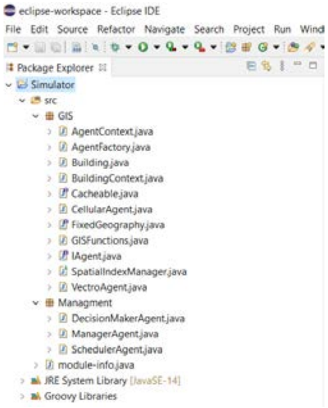


Fig. 10. Proposed Architecture.


eclipse-workspace - Eclipse IDE
File Edit Source Refactor Navigate Search Project Run Wind
Package Explorer

```
Simulator
  src
    GIS
      AgentContext.java
      AgentFactory.java
      Building.java
      BuildingContext.java
      Cacheable.java
      CellularAgent.java
      FixedGeography.java
      GISFunctions.java
      IAgent.java
      SpatialIndexManager.java
      VectroAgent.java
    Managment
      DecisionMakerAgent.java
      ManagerAgent.java
      SchedulerAgent.java
    module-info.java
  JRE System Library [JavaSE-14]
  Groovy Libraries
```

Fig. 11. Project Environment in Eclipse.

V. Experimentation and Results

The urban simulator that we are developing must provide some answers to the following questions:

- How can the urban system evolve at the municipal level?
- What is the impact of the urban growth rate on the system?
- How can an urban planning policy influence the evolution of the system?
- What is the best spatial distribution of public facilities?

To answer these questions, we went to test the simulator according to two scenarios:

A. Scenario 1

In this scenario, the decision maker - experimenter specifies the files zoning and services and chose the simulation time step (1 year, 5 years or 10 years) and then run the simulation. In this case, the simulation is based on real urban parameters of the city (population growth rate, residential sprawl ...). According to the study of the urban system of Casablanca, the municipality "Dar Bouazza" has the following real urban parameters:

- Demographic growth rate: 9.8%.
- Residential sprawl of individual housing: 24,5 Ha.
- Residential sprawl of multifamily housing: 27 Ha.
- Residential sprawl of collective housing: 38.9 Ha.
- Time step: 10 years.

These parameters are read directly from a file from the GIS, we present in Fig. 12 the map of "Dar Bouazza" at t = 0 and the result of the simulation after 10 years in Fig. 13.

By applying a 10-year simulation, we obtain the following result.

We note that with the current urbanization rate, after 10 years the consumption of space in "Dar bouaaza" would be only 31%. At the end of the simulation, approximately 3100Ha of bare land remains. With the demographic growth rate of 9.8%, the population of "Dar Bouaaza" would be of the order of 680,000 inhabitants after 10 years. This demographic increase will require the installation of new public services as indicated in Table III.

B. Scenario 2

In this scenario, the decision maker - experimenter is able to change parameters of the system: population growth rate, multi-family residential sprawl, individual residential sprawl and collective housing sprawl. The experimenter specifies the files zoning and services and chose the step of the simulation (1 year, 5 years or 10 years), set urban parameters then run the simulation.

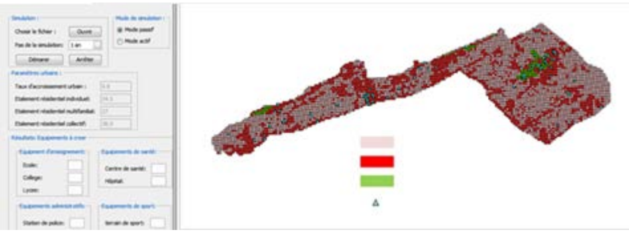

Simulation :
Choisir le fichier : Ouvrir
Pas de la simulation: 1 an
Démarrer Arrêter
Mode de simulation :
● Mode passif
○ Mode actif
Paramètres urbains :
Taux d'accroissement urbain : 9.8
Etalement résidentiel individuel: 24.5
Etalement résidentiel multifamilial: 27
Etalement résidentiel collectif: 38.9
Résultats: Equipements à creer
Equipement d'enseignement:
Ecole:
College:
Lycee:
Equipements de santé:
Centre de santé:
Hôpital:
Equipements administratifs:
Station de police:
Equipements de sport:
terrain de sport:

Fig. 12. Simulation results for Scenario 1 at t=0.

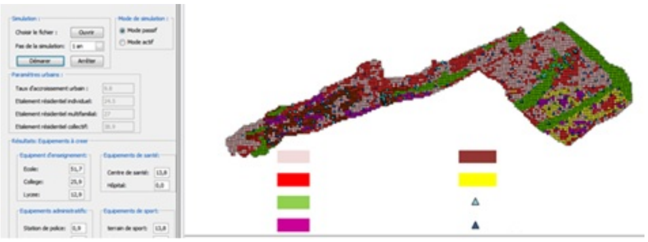
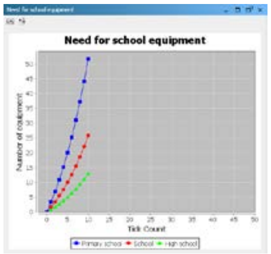
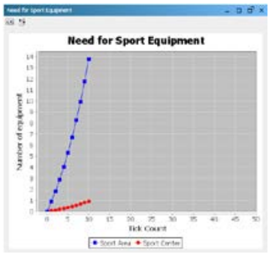
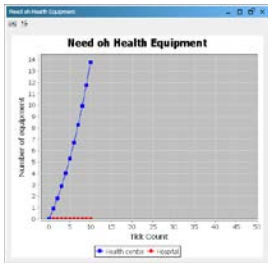
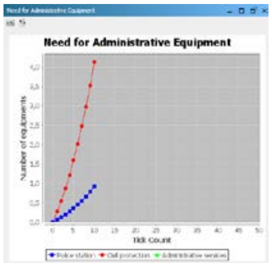
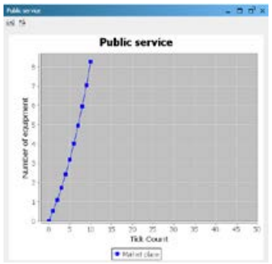
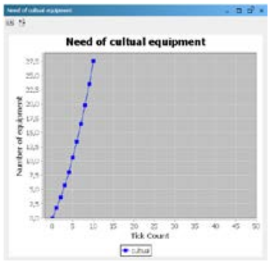

Simulation :
Choisir le fichier : Ouvrir
Pas de la simulation: 1 an
Démarrer Arrêter
Mode de simulation :
● Mode passif
○ Mode actif
Paramètres urbains :
Taux d'accroissement urbain : 9.8
Etalement résidentiel individuel: 24.5
Etalement résidentiel multifamilial: 27
Etalement résidentiel collectif: 38.9
Résultats: Equipements à creer
Equipement d'enseignement:
Ecole: 51,7
College: 25,9
Lycee: 12,9
Equipements de santé:
Centre de santé: 13,8
Hôpital: 0,0
Equipements administratifs:
Station de police: 0,9
Equipements de sport:
terrain de sport: 13,8

Fig. 13. Simulation results for Scenario 1 at t = 10.

TABLE III. SERVICES NEEDS ACCORDING TO SCENARIO 1 RESULTS

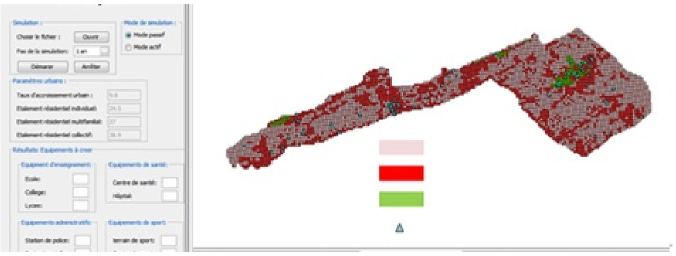
We will now test a simulation scenario where the urban planner chooses to focus "Dar Bouaaza" to be an area dedicated mainly to the multifamily housing. He chooses as a multi-family residential sprawl high compared to others and a demographic growth rate higher than the first scenario:

- Demographic growth rate: 12%.
- Residential sprawl of individual housing: 300 Ha.
- Residential sprawl of multi-family housing: 100 Ha.
- Residential sprawl of collective housing: 50 Ha.
- Time step: 10 years.

We present in Fig. 14 the initial state of the simulation at t=0. We observe in Fig. 15 that after 8 iterations (8 years) the space in "Dar Bouaaza" is completely consumed.

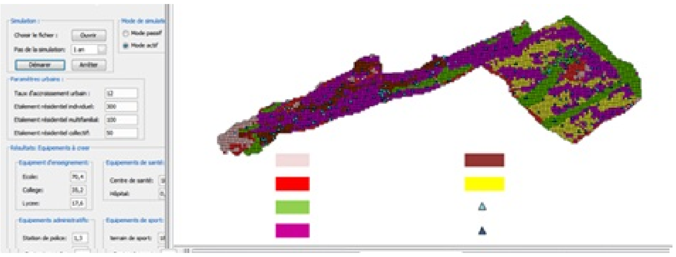
We note that after 8 iterations, the urban space of "Dar Bouaaza" is totally consumed. We conclude that with the urban parameters chosen by the decision-maker, "Dar Bouaaza" will be saturated after eight years and will no longer be able to receive new urban projects.

By applying a 1-year simulation 8 times, we obtain the following result:



Simulation :
Choisir le fichier : Ouvrir
Pas de la simulation: 1 an
Démarrer Arrêter
Mode de simulation :
● Mode passif
○ Mode actif
Paramètres urbains :
Taux d'accroissement urbain : 5.6
Etalement résidentiel individuel: 24.5
Etalement résidentiel multifamilial: 27
Etalement résidentiel collectif: 38.9
Résultats: Equipements à créer
Equipement d'enseignement:
Ecole:
College:
Lycee:
Equipements de santé:
Centre de santé:
Hôpital:
Equipements administratifs:
Station de police:
Equipements de sport:
terrain de sport:

Fig. 14. Simulation results for Scenario 2 at t=0.

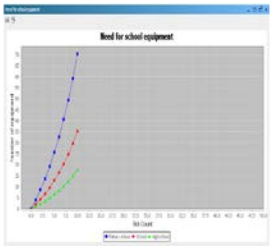
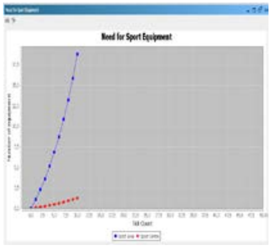
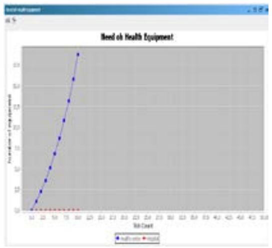
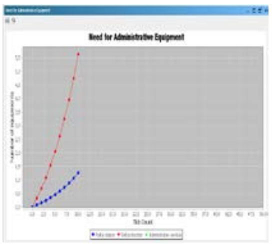
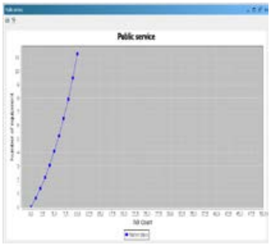
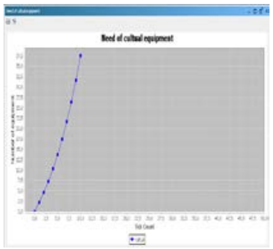


Simulation :
Choisir le fichier : Ouvrir
Pas de la simulation: 1 an
Démarrer Arrêter
Mode de simulation :
○ Mode passif
● Mode actif
Paramètres urbains :
Taux d'accroissement urbain : 12
Etalement résidentiel individuel: 300
Etalement résidentiel multifamilial: 100
Etalement résidentiel collectif: 50
Résultats: Equipements à créer
Equipement d'enseignement:
Ecole: 70,4
College: 35,2
Lycee: 17,6
Equipements de santé:
Centre de santé:
Hôpital:
Equipements administratifs:
Station de police: 1,3
Equipements de sport:
terrain de sport:

Fig. 15. Simulation results for Scenario 2 at t=8.

With the chosen demographic growth rate, there will be a significant demographic increase which will require the installation of new public services as indicated in Table IV.

TABLE IV. SERVICES NEEDS ACCORDING TO SCENARIO 2 RESULTS

C. Discussion

The results of a simulation are considered valid if firstly, the simulated habitat zoning coincides with that provided for in the development plan of “Dar Bouaaza” produced by the town planners. Secondly, if a service created is placed at a given location which is close to the position planned in the development plan.

To validate the results of our simulator, we will rely on the development plan of "Dar Bouazza". Indeed, this plan has been developed based on certain assumptions which are: the municipality of "Dar Bouazza" is a peripheral urbanized area on which the balance of the functional structure of the city Casablanca. It’s an urban pole that will guarantee urbanization of a part of demographic growth of other municipalities.

According to the development plan, we will simulate the municipality "Dar Bouazza" using the following parameters:

- Step time: 10 years.
- Demographic growth rate: 4.3%.
- Residential sprawl of individual housing: 165.2 Ha.
- Residential sprawl of multifamily housing: 118.8 Ha.
- Residential sprawl of collective housing: 78, 4 Ha.

To evaluate the results of the simulator, we do a geographical intersection between the exported layer from the simulator and the development plan layer to identify areas of change. Results are presented in Fig. 16.

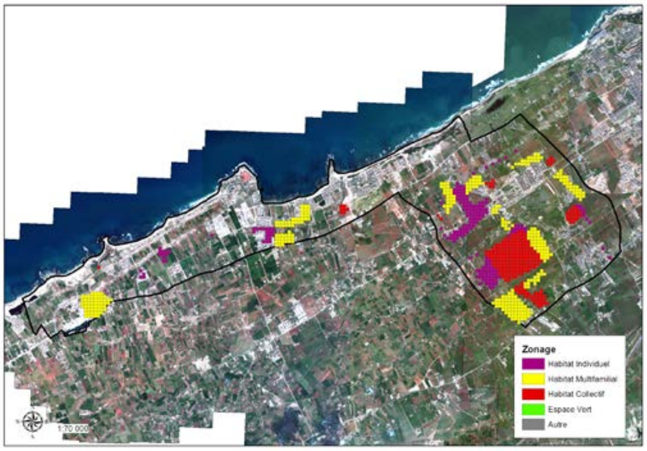


Fig. 16. Area of mismatch between the Management Plan and the Simulation results.

In Table V, we present the surface areas and percentages change between the simulator results and the development plan produced by urban planners.

We conclude that the simulator has a percentage of accuracy about 80% for the simulation of residential areas.

To evaluate the results of the simulation of services, we chose to test an example of educational service which is schools. The development plan of "Dar Bouaaza" provided the creation of fifty schools across the territory of the municipality.

TABLE V. EVALUATION OF HOUSING RESULTS

Zone	Number of unmatched cells	Total number of cells	Error rate
Individual housing	204	1652	12%
Multi-familial housing	303	1188	25%
Collective housing	240	784	30%
Total	**747**	**3624**	**20%**

In the context of our simulator, vector agent "service" of school-type is created in accordance with a set of rules which are: a school must be created in a residential area, a school must be created away from any source of pollution (industry, waste disposal ...), a school has to be created close to some attractive equipment (library, theatre, ...), a school must be served by road with a width that does not exceed 40m, the maximum distance between two schools should not exceed 500m because a child cannot walk more than 500m, etc.

After an interaction between the different agents of the system, the "Service" agent finds the best location. To evaluate the spatial distribution of the results, we applied a buffer zone on simulated schools and schools planned by the development plan of “DarBouaaza” and then we identified the points that don’t coincide, the result is presented in Fig. 17.

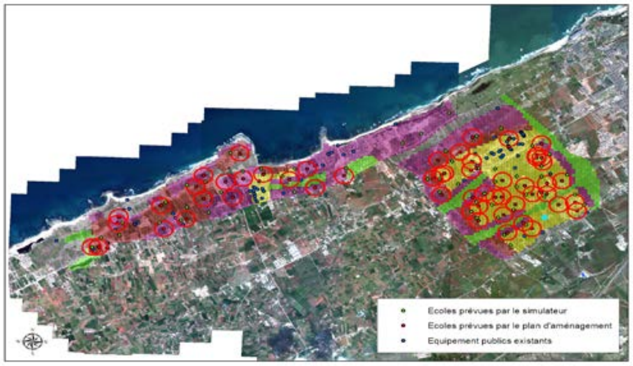

● Ecoles prévues par le simulateur
● Ecoles prévues par le plan d'aménagement
● Equipement publics existants

Fig. 17. Buffer Zone applied to Services Planned by the Development Plan of “Dar Bouaaza”.

The schools planned by the simulator that do not coincide with those in the development plan are presented in Fig. 18:

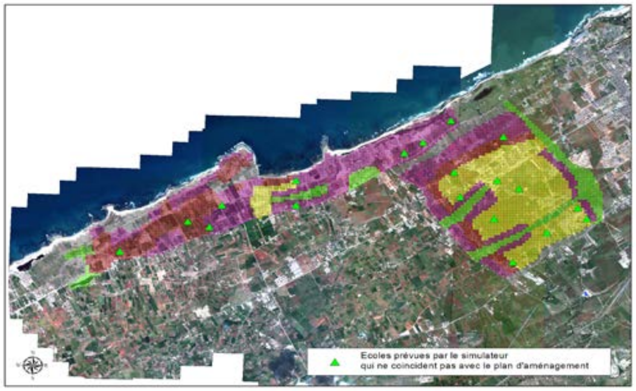

▲ Ecoles prévues par le simulateur qui ne coincident pas avec le plan d'aménagement

Fig. 18. Schools Planned by the Simulator which do not Coincide with the Development Plan of “Dar Bouaaza”.

Among 55 schools to be created, there are 19 schools that have a spatial position that does not coincide with the schools planned in the development plan of "Dar Bouaaza", which gives a percentage of accuracy of 65% simulator.

VI. CONCLUSION

The simulation of urban system is a broad field that can be approached from several perspectives. In our research, we focused on the characteristics of the agent paradigm to simulate these systems. Thus, we have proposed a new approach based on the use of cellular agents and vectors agents to represent the spatial area. We tested our proposed model with a case study of “Dar Bouaaza” municipality. We run two simulation scenarios by changing in each scenario the hypotheses concerning urban planning especially in terms of demographic growth rate and residential sprawl. However, the model can be extended to simulate other urban planning aspects such as the infrastructure, industry zoning and natural spaces.

REFERENCES

[1] K. Torres , A. Torres, L. Pietrzyk, J. Lisiecka, M. Błoński, M. Bącik-Donica, G. Staśkiewicz, R. Maciejewski, “Simulation techniques in the anatomy curriculum: review of literature” Folia Morphol (Warsz). 2014 Feb;73(1):1-6. doi: 10.5603/FM.2014.0001. PMID: 24590516.

[2] A. Langlois, C.Phipps, “Automates cellulaires, Application à la simulation urbaine”. Hermès, 1997.

[3] D. O’Sullivan , P.M. Torrens “Cellular Models of Urban Systems” In: S. Bandini, T. Worsch (eds) Theory and Practical Issues on Cellular Automata. 2001 Springer, London. https://doi.org/10.1007/978-1-4471-0709-5_13.

[4] J. Garrido “ Models and Simulation”. In: Object Oriented Simulation. Springer, Boston, MA. 2009. https://doi.org/10.1007/978-1-4419-0516-1_1.

[5] J. Ferber, “Agent-based simulation”, LIRMM - Université Montpellier II, (2009).

[6] F. Bousquet, C. Le Page, “Multi-agent simulations and ecosystem management: a review”, Elsevier, (2004).

[7] John B. Davis, “Agent-Based Modeling’s Open Methodology Approach: Simulation, Reflexivity, and Abduction”, Marquette University and University of Amsterdam, pp 509-529. 2018. https://doi.org/10.4000/oeconomia.4402.

[8] A. Rodrigues, J. Raper, “Defining spatial agents”. In: Spatial Multimedia and Virtual Reality Research Monographs Series, Eds J Raper, A Câmara (Taylor and Francis, London) 1999 pp. 111 – 129.

[9] Y. Hammam, A. Moore, P. Whigham, C. Freeman, “A vector-agent paradigm for dynamic urban modelling”. In: Proceedings of 15th Annual Colloquium of the Spatial Information, Research Centre, Dunedin, New Zealand, 2003.

[10] J. Ferber “Les Systèmes Multi-Agents, vers une intelligence collective ” 1995 InterEditions.

[11] Y. Demazeau “From Interactions to Collective Behaviour in Agent-Based Systems”, Proceedings of the First European Conférence on Cognitive Science, Saint-Malo 1995.

[12] J. Ferber “Agent-based simulation”, LIRMM - Université Montpellier II, 2009.

[13] S. Albeverio, D. Andrey, P. Giordano, A. Vancheri, "The dynamics of complex urban systems: An interdisciplinary approach", Berlin, Springer, 2008.

[14] M. Batty "Cities and Complexity: Understanding Cities with Cellular Automata, Agent-Based Models, and Fractals", MIT Press, Cambridge, 2005.

[15] F. Semboloni "From spatially explicit to multiagents simulation of urban dynamic", Advances in Complex Systems, Volume 10, Issue. 2, 2007, Pages 355-362.

[16] I. Benenson, P. Torrens "Geosimulation: Automata-Based Modeling of Urban Phenomena", John Wiley & So, 2004.

[17] D. Fecht, L. Beale, D. Briggs, "A GIS-based urban simulation model for environmental health analysis", Environmental Modelling & Software, Vol. 58, pp. 1-11, 2014. https://doi.org/10.1016/j.envsoft.2014.03.013.

[18] P. Waddell "UrbanSim: Modeling Urban Development for Land Use, Transportation and Environmental Planning", journal of the American Planning Association 68(3): 297-314, 2002.

[19] E. J. Miller, J. Douglas Hunt, J. E Abraham, P. A Salvini,"Microsimulating Urban Systems" Computers, Environment and Urban Systems, Volume 28, Issues 1–2, 2004, pp. 9-44.

[20] T. Patrick, B. Arnaud, D. Alexis and G. Benoit and M. Nicolas and T. Chi Quang "Simulating Urban Growth with Raster and Vector models: A case study for the city of Can Tho, Vietnam". In: International conference Autonomous Agents and Multiagent Systems (AAMAS 2016), 9 May 2016 - 10 May 2016 (Singapore, Singapore).

[21] Langlois A, Phipps C (1997) Automates cellulaires, Application à la simulation urbaine. Paris, Hermès, 197 p.

[22] B. Jiang, H.R. Gimblett "An agent-based approach to Environmental and Urban systems within Geographic Information Systems". In : H. R. Gimblett (Ed.) Integrating Geographic Information Systems and Agent-Based Technologies for Modelling and Simulating Social and Ecological Phenomenon, New York, Oxford University Press, pp. 171-190, 2002.

[23] T. Berger "Agent-based spatial models applied to agriculture: a simulation tool for technology diffusion, resource use changes and policy analysis". Agricultural Economics, 25 (2-3):245-260, 2001.

[24] A. Mustafa, M. Cools, I. Saadi, J. Teller, "Coupling agent-based, cellular automata and logistic regression into a hybrid urban expansion model"(HUEM), Land Use Policy, Vol. 69, pp. 529-540, 2017. https://doi.org/10.1016/j.landusepol.2017.10.009.

[25] A. Rodrigues, C. Grueau, J. Raper, N. Neves "Environmental planning using spatial agents". In: S. Carver (ed.) Innovation in GIS 5, Taylor & Francis, London.1998.

[26] T. Arentze, H.J.P. Timmermans, D. Janssens, G. Wets G "Modeling short-term dynamics in activitytravel patterns: from Aurora to Feathers". In: Proceedings of the Innovations in Travel Modeling Conference, Austin, US, May 21–23, 2006.

[27] I. Benenson, P. Torrens "Geosimulation: Automata-based modeling of urban phenomena". John Wiley & Sons Ltd., Chichester, England. 2004

[28] L. Caneparo, F. Guerra, E. Masala "UrbanLab – Generative platform for urban and regional design". In Proceedings: 8th International DDSS Conference. Progress in Design & Decision Support Systems in Architecture and Urban Planning. Eindhoven University of Technology, The Netherlands. Jos P. van Leeuwen & Harry J.P. Timmermans Ed. 2006.

[29] D. Vanbergue, A. Drogoul "Approche multi-agent pour la simulation urbaine". Actes des 6èmes Journées CASSINI, pages 95-112. 2002.

[30] S. Maneerat, E. Daudé, "A spatial agent-based simulation model of the dengue vector Aedes aegypti to explore its population dynamics in urban areas", Ecological Modelling, Vol 333, pp. 66-78, 2016. https://doi.org/10.1016/j.ecolmodel.2016.04.012.

[31] Y. Hammam, A. Moore, P. Whigham, C. Freeman "A vector-agent paradigm for dynamic urban modelling". In: Proceedings of 15th Annual Colloquium of the Spatial Information, Research Centre, Dunedin, New Zealand. 2003.

[32] Y. Hammam, A. Moore A, P. Whigha (2007) "The dynamic geometry vector agents". Computers, Environment and Urban Systems 31(5), pp. 502-519. 2007.

[33] T. Patrick, « La modélisation du temps dans la simulation à base d'agents », L'Information géographique, Vol. 79, p. 65-78. 2015. DOI : 10.3917/lig.792.0065.

[34] S. Lamy, A. Ruas, Y. Demazeau, C. Baeijs, M. Jackson, W. Mackaness, R. Weibel "AGENT Project: Automated Generalisation New Technology". 5th EC-GIS Workshop, Stresa(I), 28-30 June 1999, in Proceeding, pp.407-415, Edited by K. Fullerton, laboratoire COGIT, IGN-SR 990035/S-COM, 1999.

[35] A. Ruas "The role of meso objects for generalisation". 9th International Symposium on Spatial Data Handling (SDH'00), 10-12 August, 2000. Beijing (China).

[36] C. Duchêne, M. Barrault, N. Regnaud, K. Haire, C. Baeijs, P. Hardy, W. Mackaness, A. Ruas, R. Weibel (2001) "Integrating multi-agent, object oriented alogorithm techniques for improved automated map generalization", 20th International Cartographic conference (ICC'01), 6-10 August, 2001. Beijing (China).

[37] F. Barramou, M. Addou, « An agent based approach for simulating complex systems with spatial dynamics application in the land use planning", in International Journal of Advanced Computer Science and Applications(IJACSA), Vol.3 Issue 11, 2012.

[38] C. Duchêne, N. Regnauld "Le modèle AGENT". In : Ruas A., Généralisation et représentation multiple, Paris, Hermes Lavoisier, Chapitre 21, pp 369-386.2002.

[39] Repast Simphony "Repast Documentation". Available online at : http://repast.sourceforge.net/docs.php (accessed on November 2020)

[40] ArcGIS, "Technical Support". Available online at : www.esri.com (accessed on November 2020).

Applying Digital Image Processing Technology in Discovering Green Patches in the Desert of Saudi Arabia

Ali Mehdi[1], Md Alamin Bhuiyan[2]
Department of Computer Engineering, College of Computer Sciences and Information Technology
King Faisal Universitym, P.O. Box 380, Al-Ahsa 31982, Saudi Arabia

Abstract**—In recent years, the Kingdom of Saudi Arabia has witnessed a noticeable growth of grass and small trees in the desert, forming green patches. Those green patches may have the potential to spread and cover a wider area in the desert in the coming years, thus, making areas of the desert potential agricultural land. This research aims to detect the change of green patches in the desert of Saudi Arabia to solve the challenge that is mainly due to the lack of an organized dataset. Using a series of satellite images of the desert landscape, a change detection algorithm is used to identify the changes in green spaces. This algorithm includes the presentation of multi-temporal datasets to evaluate the chronological special effects. This paper presents an optical flow analysis among images captured at different time sequences. The algorithm shows promising results of change detection in green patches in the desert of Saudi Arabia detected by color segmentation. The algorithm has been validated over a set of satellite images demonstrating an effective performance.**

Keywords—Image processing; change detection; optical flow analysis; green patches; color segmentation

I. Introduction

In recent years, The Kingdom of Saudi Arabia has taken considerable strides to achieve food security, a goal that is one of the noble goals of Vision 2030. One of the strategies in achieving food security is to increase the agricultural area within the 2.15 million km^2 (approx.) of the total area of the Kingdom of Saudi Arabia [1]. Although desert covers most of the Saudi Arabian land, historical images of Google Earth show a spread of green patches over the years. This observation was also reported by NASA in 2012 [2]. Some of those green patches are natural and others are artificial. The rationale behind this research is to contribute towards achieving food security in Saudi Arabia through detecting the spread of green patches using Change Detection (CD) algorithm. CD is the method of discovering variations or changes that occur in a series of images over a period of time. CD includes the presentation of datasets based on time series, thus, evaluating the evaluating the unique chronological properties. For the previous decades, Remote Sensing Data, for example, Thematic Mapper (TM) [3-5] and Satellite Probatoire d'Observation de la Terre (SPOT) [6], were the foremost sources of data for a number of CD applications. This is due to the benefits of: (1) data attainment that is performed repeatedly, (2) the synoptical assessment, and (3) the ability for computers to process digital data.

In this research, the proposed algorithm is implemented for sequences of satellite images for optical flow analysis. The images' capture, ranges from 1990 to 2020 depending on the region and on the available satellite records. The results exhibit a promising outcome of green patches detected using color segmentation.

The rest of this paper is organized as the following: Section 2 discussed some related work, particularly, in landscape analysis of green patches and the implementation of CD algorithms. The detailed implementation of the algorithm used in this paper is described in Section 3. The results are presented in Section 4, and finally this paper is concluded in Section 5.

II. Related Work

A. Landscape Analysis of Green Patches

The author in [7] used the city of Hangzhou in China to validate their methodology, which maps landscape evaluation with the sub-pixel in order to investigating the spatiotemporal sequence that might emerge. Their methodology also examines the changes of classified green patches of urban areas. To deriving the area of green patches at the level of sub-pixel, the method examines (more than once) the spectral mixture that was used within the dataset of time series Landsat. Analyzing the landscape measure was performed to distinguishing the various green patches sequences of urban areas. Based on their methodology, a green space network was shaped, aiming to increase the green patches in the urban areas of Hangzhou. According to the authors, their study results should assist in refining the perceptive of the growth of assorted urban areas.

The author in [8] proposed an automated algorithm called Density Weighted Connectivity of Grass Pixels (DWCGP) in order to evaluate the roadside grass biomass from images to identify fire-prone regions. This method computes the vertical pixels length of connected grass in each column of a given image. After this step, the algorithm weights the length using the density of the grass in the column's region. Pixels classification is then conducted by utilizing Feedforward Artificial Neural Networks, whereas the main feature alignment at each pixel is calculated using multi-orientation Gabor Wavelet Filter vote. Their algorithm works well with

This research was sponsored by the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia.

grass stems that are not vertical, also it adapts to variations of Gabor filter parameters and surrounding region widths.

A fuzzy logic approach was conducted using Hadoop Hive [9]. This approach discovers the harvests plantation information from the agro-climatic historical database during the period of 1983-2016 taking Egypt as the test ground. Their approach gives a number of circumstances for the plantation dates for each harvest with a suitability level for every situation. Their proposed approach should help managing the procedure of harvests plantation, taking into consideration factors like: harvesting dates, diseases and watering requirements data.

In a contrary work [10], the authors conducted a study that evaluates the variations of drought in the Tihama Plain in Yemen that contributes towards over 40% of agricultural products in the country. This study considers the time span and its in relation to the drought spread. In their work, the Standardized Precipitation Index (SPI) was implemented in order to assess the status of drought temporally. Geographic Information Systems (GIS) was also implemented to display the different spreading of drought spatially in the region of interest (study area). Their results show a concerning drought phenomenon over three decades of time.

The author in [11] investigated the green field detection methods for visible satellite images using the vegetative index (COM) in addition to the development of two methods to increase detection accuracies, which are enhanced versions of COM. The visible satellite images included six locations and four zoom levels obtained through Google Map. The three methods were applied in their experiment in order to compare the accuracies in detection.

B. Implementation of Change Detection Algorithms

The author in [12] proposed an algorithm to generate the Landsat Burned Area Essential Climate Variable (BAECV) by recognizing burned spaces in dense time-series of Landsat data. Gradient boosted regression models were utilized in the generation of possible burn spaces by implementing band values and spectral indices from separate Landsat views, insulated surroundings, and varied measurements between the view and insulated surroundings. By utilizing pixel-level thresholding in conjunction with an area increasing procedure, the categorizations of burn were produced from the burn possibility surfaces. 1984 to 2015 was the time duration in which the BAECV results were produced for, using the United States (US) as the test ground. For each Landsat scene, the BAECV products were comprised of pixel-level burn probabilities, as well as annual combination that include the maximum burn probability and a burn organization.

The author in [13] proposed a collective of time series algorithms aiming to improve the observation of land change. Their methodology combines the algorithms of Continuous CD and Classification (CCDC) [14] and Cumulative Sum of Residuals (CUSUM) for break detection; for removing false positives (or breaks in time spans not demonstrating land change), the Chow Test [15] was utilized. The combination of those three algorithms, were implemented to three Landsat views taking the US as the test ground. The results were measured according to their capability in properly distinguish structural breaks from steady time spans.

The work presented in [16] display some initial results for CD by utilizing Landsat and Worldview images. The area of interest that the research targeted had some substantial development between May 2014 and October 2015 like buildings emerging. The authors investigated several methodologies for CD. The authors implemented pansharpening to improve the resolution to fifteen meters, and that was related to Landsat images. A chronochrome covariance equalization was also implemented between two Landsat images. The authors then analysed the residual between the two equalized images by utilizing various algorithms, for example, direct subtraction and global Reed-Xiaoli (GRX) detector.

The author in [17] applied images for Synthetic Aperture Radar (SAR) in order to observe disasters and for the purpose of environmental monitoring. The aim of their work is to minimize the effect of noise on SAR image CD. They built their work on mathematical morphology filtering and K-means clustering for SAR image CD. As an initial step, the authors used logarithmic transformation to transform multiplicative noise into additive noise in two SAR images. Then morphological filtering was used to de-noise the two multitemporal SAR images. Finally, in order to find the two distinctive images, the authors used the mean ratio operator and subtraction operator.

The author in [18] have proposed a Time-Series Classification approach based on CD (TSCCD) for rapid Land Use/Land Cover (LULC) mapping, which utilizes the Prophet algorithm in order to find the ground cover change points, then implement time sequence dissection in a time component and the Dynamic Time Warping (DTW) algorithm in order to categorize the sub-time sequence. Time sequence images of the City of Wuhan were selected for testing during the period of 2000 to 2018. For validation, the authors utilized China’s national land-use surveys for the years: 2000, 2005, 2008, 2010, 2013, and 2015. In their research, they assumed that the ground cover stays unaffected in each subsequence, only one time-training sample selection and one LULC categorization were required in order to improve the effectiveness of their work. According to the authors, the Prophet can spot large and slight variations to a precise degree, capture the direction and rate of change, as well as its suitability to process noise and missing data. Because data is subject to irregular observations or missing values, DTW is primarily utilized to enhance the accuracy of time sequence categorization and to fix the time misalignment issues of ground cover series data.

Another algorithm was proposed in the CD field taking the urban renewal in the City of Guangzhou (China) as a test ground for the period between 2000 and 2010 [19]. Their research suggests a technique to extract and map yearly Impervious Surface Percentage (ISP) and to differentiate patterns in urban growth by utilizing time sequence images of average resolution. Their research was carried on by applying the Cubist tree model for annual ISP inversion (AoCubist), enhancing multi-temporal Landsat merged images to reduce the influence of phenology and inter-year climate change. They also developed the C5.0 decision tree algorithm with temporal-spatial filtering guidelines in order to enhance the space-time endurance and the separation of patterns derived by unsupervised K-means categorization. Their results were validated against Google Earth images, which indicate that their categorization achieved an overall accuracy of 88.32% to 90.85%. Whereas, the yearly urban growth rate stayed between 4% and 10%, while the yearly deurbanization rate ranged between 1% and 5%.

The proposed methodology described in the next section differs from those discussed above. It is a combination of algorithms that are applied to detect green patches in the desert of Saudi Arabia using satellite images.

III. Methodology

Three CD algorithms are widely used to detect the changes in green spaces: (1) image differencing, (2) image ratioing, and (3) Principal Component Analysis (PCA) [20]. Image differencing is accomplished by taking the absolute value of the subtraction of the second image from the first image. Image ratioing is accomplished by calculating the ratio of the amount of green space covered in two registered images captured at different times.

However, this apparent intensity change at each pixel is due to various factors like camera movement, change in position and intensity of light sources, object movement, deformation, specular reflections, and so on.

Although the image differencing is a simple method of CD, but it produces only the magnitudes of the changes. Ratioing is unable to describe changes efficiently in the region of interest (the study area). It produces partial change information and provides the least accuracy in comparison to other methods, thus in plane areas, it is discouraged. PCA approaches are employed for CD in sort of vegetation and water feature.

This research presents an optical flow analysis among images captured at different time sequences. Optical flow is given as the movement of discrete pixels within the area of an image. Optical flow usually gives a reasonable estimation of the actual physical motion applied within the area of an image [21].

The main idea of optical flow is derived from Lucas and Kanade [22], who presented an iterative registration technique for stereo employing motion detection called the local method of optical flow. Horn and Schunk [23] measured the optical flow using a global method which offers the best density in the optical flow. The popular optical flow spatiotemporal method of Black and Anandan [24] was one of the methods that were based on [22] and [23]. The author in [25] presented a combined global and local optic flow techniques which conglomerates the benefits of global and local methods. This research presents a Local Global Hybrid (LGH) method which employs the Lucas and Kanade's least-square fit approach a prototype of the local method, whereas the Horn and Schunck approach was utilized as a characteristic for a global method.

A. Optical Flow Analysis

Let $\{f_1, f_2, ..., f_N\}$ be a set of N image sequences received at time $\{t_1, t_2, ..., t_N\}$ over satellite imagery. Every image maps a two-dimensional pixel coordinate $(x, y) \in R^k$ representing the brightness or color (k=1 for gray-scale images and k=3 for RGB color images). A motion vector is determined by considering a correspondence between rectangles at times $\{t_1, t_2, ..., t_N\}$, where t is the time index in image sequences:

$$m_v = \arg\min \left\| f_1(x, y, t) - f_2(x-a, y-b, t+1) \right\| \tag{1}$$

Most of the researcher use block-matching algorithm to determine the motion vector. Nevertheless, the main drawback of using block-matching algorithm for determining the motion vector is that its computational cost is high. N^2 different possible velocity vectors have to be checked in the image, each of which requires N^2 additions for every block. Numerous sub-optimal search strategies try to obtain the approximate velocity vector, without really evaluating all $\boldsymbol{N}^2$ of the possible velocity vectors. Optical flow deliberates the motion vector as a function of continuous position, rather than discrete position.

The Horn–Schunck technique of approximating optical flow is a global approach which presents a global restriction of smoothness. The flow is expressed by a global energy function that is required to be reduced:

$$E = \iint [(f_x a + f_y b + f_t)^2 + \alpha^2 (|\nabla a|^2 + |\nabla b|^2)] dxdy \tag{2}$$

where f_x, f_y and f_t are the derivatives of the image intensity values along the x, y and time dimensions respectively. The optical flow vector is $\mathbf{v} = [a, b]^T$, and the parameter α (>0) is a regularization constant.

This research outlines a Local–Global Hybrid (LGH) approach for optical flow analysis with the following notations:

$$w = (a, b, t)^T \tag{3}$$

$$|\nabla w|^2 = |\nabla a|^2 + |\nabla b|^2 \tag{4}$$

$$\nabla_3 f = (f_x, f_y, f_z)^T \tag{5}$$

$$J\rho(\nabla_3 f) = k\rho * (\nabla_3 f \ \nabla_3 f^T) \tag{6}$$

Obviously, the Lucas–Kanade approach reduces the quadratic form given by the following equation:

$$J\rho(\nabla_3 f) = k\rho * (\nabla_3 f \ \nabla_3 f^T) \tag{7}$$

The Horn–Schunck approach reduces the function given below:

$$E_{HS}(w) = \int_{\Omega} \left\lfloor w^T J_0(\nabla_3 f) w + \alpha (|\nabla w|^2) \right\rfloor dxdy \tag{8}$$

This expression recommends a likely technique to spread out the Horn–Schunck equation to the required LGH method. Once the matrix $J_0(\nabla_3 f)$ is replaced by the structure tensor $J_\rho(\nabla_3 f)$ with selected integration scale $\rho > 0$, (9) is established.

$$E_{LGH}(w) = \int_\Omega \left\lfloor w^T J_\rho(\nabla_3 f)w + \alpha(|\nabla w|^2)\right\rfloor dxdy \tag{9}$$

The reduced flow field (*a*, *b*) fulfills the Euler– Lagrange equations expressed by the following equations:

$$0 = \Delta a - \frac{1}{\alpha}(K_\rho * f_x^2 a + K_\rho * f_x f_y b + K_\rho * f_x f_t) \tag{10}$$

$$0 = \Delta b - \frac{1}{\alpha}(K_\rho * f_y^2 b + K_\rho * f_x f_y a + K_\rho * f_x f_t) \tag{11}$$

The aforementioned methods applied were merely the spatial smoothness functions. However, the spatiotemporal expressions provide more suitable results than spatial expressions due to the extra de-noising belongings in the temporal direction [26]. A spatiotemporal alternate of the Lucas–Kanade method just substitutes the convolution of two-dimensional Gaussians by spatiotemporal convolution of three-dimensional Gaussians. This leads to a 2-by-2 linear system of equations for a and b that are unknown. Combining the Lucas–Kanade's temporal extended variant method with the Horn–Schunck method, spatiotemporal version of the proposed LGH function is established (12) [26].

$$E_{LGH3}(w) = \int_{\Omega\times[0,T]} \left\lfloor w^T J_\rho(\nabla_3 f)w + \alpha(|\nabla w|^2)\right\rfloor dxdy \tag{12}$$

where convolutions with Gaussians are performed in a spatiotemporal approach and it is given below:

$$|\nabla_3 w|^2 = |\nabla_3 a|^2 + |\nabla_3 b|^2 \tag{13}$$

If J_{pq} is the component (*p*,*q*) of the structure tensor, $J_\rho(\nabla_3 f)$, the Euler–Lagrange equations are provided by the equations below:

$$0 = \nabla_3 a - \frac{1}{\alpha}(J_{11}a + J_{12}b + J_{13}) \tag{14}$$

$$0 = \nabla_3 b - \frac{1}{\alpha}(J_{21}a + J_{22}b + J_{23}) \tag{15}$$

If the spatiotemporal Laplacian substitutes the spatial Laplacean, then the following equation will be achieved.

$$\Delta_3 = \partial_{xx} + \partial_{yy} + \partial_{11}. \tag{16}$$

B. Proposed Algorithm for LGH

For a rectangular pixel grid, functions $a(x,y,t)$ and $b(x,y,t)$ are unknown, whereas the *a* estimation at some pixel *k* with (*k*= 1,. . . ,*N*) can be considered as a_k. If J_{pqk} is the component (*p*, *q*) of the tensor $J_\rho(\nabla f)$ of pixel *k* and *N*(*k*) represents the set of (4 in 2D, 6 in 3D) neighbours of pixel *k*, so a finite difference estimation to the Euler–Lagrange equations are provided by the following:

$$0 = \sum_{j\in N(k)} \frac{a_j - a_k}{h^2} - \frac{1}{\alpha}(J_{11k}a_k + J_{12k}b_k + J_{13k}) \tag{17}$$

$$0 = \sum_{j\in N(k)} \frac{b_j - b_k}{h^2} - \frac{1}{\alpha}(J_{21k}a_k + J_{22k}b_k + J_{23k}) \tag{18}$$

where *k* = 1,. . . , *N*. This sparse linear system of expressions has solved the Successive Over-Relaxation (SOR) methodology due to its simplicity and effectiveness [26]. When the upper index represents the iteration step, (18) and (19) can express the SOR method.

$$a_k^{i+1} = (1-\omega)a_k^i + \omega\frac{\sum_{j\in N^-(k)} a_j^{i+2} + \sum_{j\in N^+(k)} a_j^i)}{|N(k)| + \frac{h^2}{\alpha}J_{11k}} - \omega\frac{\frac{h^2}{\alpha}(J_{12k}b_k^i + J_{13k}}{|N(k)| + \frac{h^2}{\alpha}J_{11k}} \tag{19}$$

$$b_k^{i+1} = (1-\omega)b_k^i + \omega\frac{\sum_{j\in N^-(k)} b_j^{i+2} + \sum_{j\in N^+(k)} b_j^i)}{|N(k)| + \frac{h^2}{\alpha}J_{22k}} - \omega\frac{-\frac{h^2}{\alpha}(J_{22k}a_k^i + J_{23k}}{|N(k)| + \frac{h^2}{\alpha}J_{22k}} \tag{20}$$

where

$$N^-(k) = \{k \in N(k), j < k\}, \quad N^+(k) = \{k \in N(k), j > k\}.$$

Here $|N(k)|$ represents the number of pixels that are adjacent to pixel *k*, which belong to the image area. The relaxation parameter $\omega \in [0,2]$ has a solid impact on the convergence speed. For $\omega = 1$ it matches exactly with the Gauss–Seidel method. Normally, the typical values for ω are [1.9, 1.99]. Considering the first iteration, the flow components are initiated by 0. The explicit option of the initialization is not vital, this is because the method is globally convergent. The algorithm is applied for a number of image sequences with time frames $\{t_1, t_2, ..., t_N\}$.

C. Optical Flow Field Construction Algorithm

Input: A set of images at times $\{t_1, t_2, ..., t_N\}$, iteration i for SOR, α , w and ρ .

Output: Optical Flow Field

Method: The Optical Flow Field is created by these steps.

```
1. Read files in matrix array.
2. For each image received at time t, perform the following steps:
      i.   Pre-smooth the image by implementing the Gaussian Kernel with standard deviation $\alpha$ .
      ii.  Compute the variation of image $f_x, f_y, f_t$ .
      iii. Calculate the velocity $a$, $b$ of image $t$ by using the (SOR) methodology for $k$ iteration.
      iv.  Compute the resulting velocity and angle, then draw.
      v.   Store the outcome.
3. End
```

IV. Experimental Results

The effectiveness of the algorithm has been validated over a set of satellite images. The images were collected from Google Earth. A sample image of the study area is shown in Fig. 1.

Experiments have been conducted using Visual C++ software in the Windows 10 operating system. The computer's setup was Intel® Core™ i5 CPU with 2.70 GHz, and 2 GB RAM.

The classified map of the study area has been constructed, illustrating the surface cover using ArcGIS. Fig. 2 illustrates the dominant surface cover of the study area. The green spaces are presented in green colors, orange presents grounds, sand dunes are presented in yellow, buildings are presented in red, while the water bodies are presented in blue.

The outcome of the optical flow field generated by utilizing LGH methodology is presented in Fig. 3. This represents the set of (4 in 2D, 6 in 3D) neighbours of pixel that are adjacent to pixel k, which belong to the image area. This was applied to a number of image sequences with time frames.

Obviously, too much regularization reduces the flow of the changes. Optical flow outcomes better results when the regularization parameter α is from 0.05 to 0.09 for 200 to 1000 iterations. The green areas are finally detected by color segmentation in *CIE a*b** color model, as shown in Fig. 4.

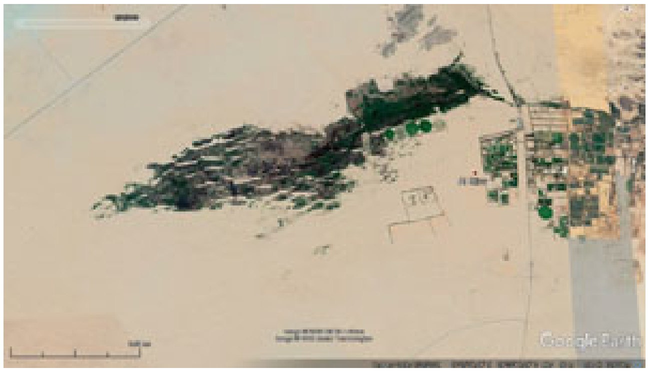


Fig. 1. A Sample Image of the Study Area.

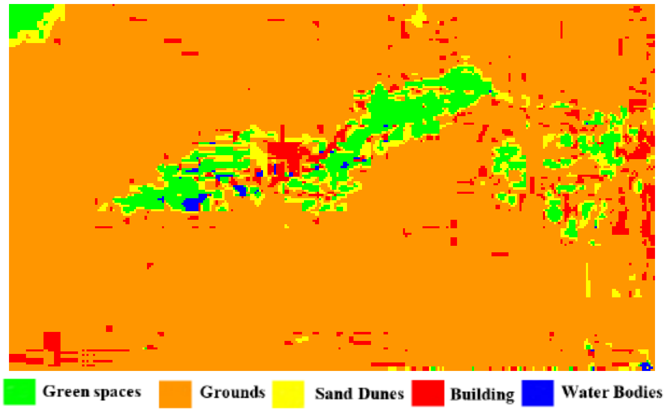


Fig. 2. A Sample Image Illustrating the Surface Cover.

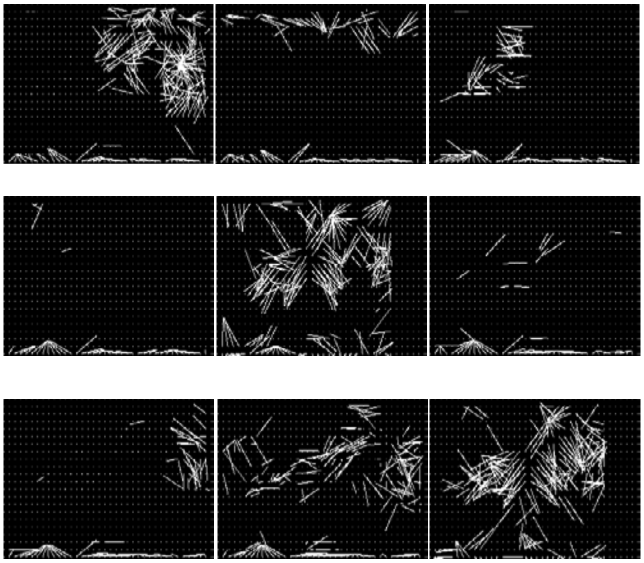


Fig. 3. Optical Flow Field Generated by utilizing LGH Methodology.

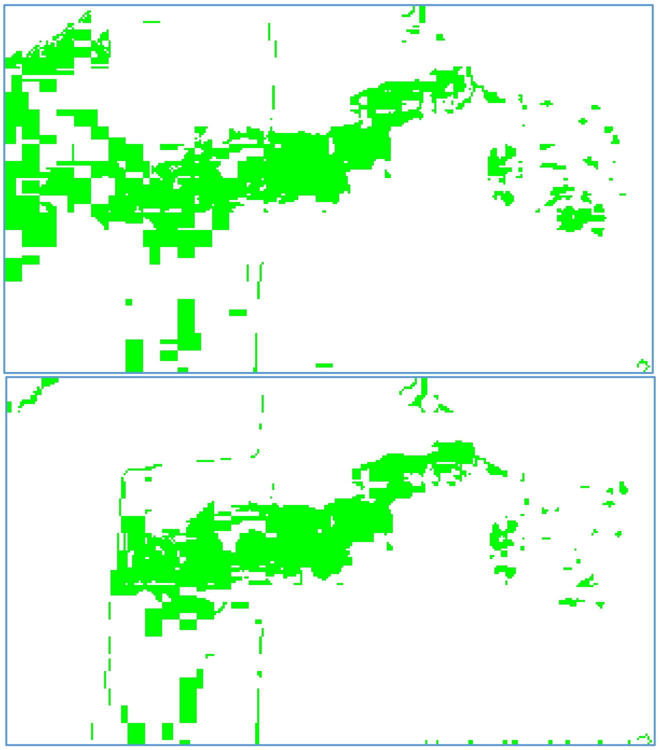


Fig. 4. Extraction of the Green Spaces from Surface Cover Images.

V. Conclusion

In order to detect the green patches in the desert of Saudi Arabia, a series of satellite images were captured that cover 2 to 3 decades of time depending on the location and on the available satellite images. CD algorithm was utilized to identify the changes in green spaces.

Optical flow analysis was implemented to the image sequences received at a certain time through the satellite imagery. Each image maps a 2D pixel coordinate representing the gray-scale or the RGB color images. A motion vector is determined by considering a correspondence between rectangles at give times.

This research outlines a Local–Global Hybrid (LGH) approach for optical flow analysis in a way that combines the temporal extended variant of both the Lucas–Kanade and the Horn–Schunck methods. The optical flow field was constructed using a set of images at given times and processed in a number of steps. The experimental results show an effective detection of the green patches from the desert area. This research will be conducted further in order to compute the rate of growth in the green patches of the Saudi desert, thus, observing a pattern that may emerge. This will then lead to further analyses and prediction of future spread of the green patches.

Acknowledgment

The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number IFT20164.

References

[1] S. Fiaz, M. A. Noor, and F. O. Aldosri, “Achieving food security in the Kingdom of Saudi Arabia through innovation: Potential role of agricultural extension”. Journal of the Saudi Society of Agricultural Sciences, vol. 17, pp. 365-375, October 2018.

[2] “NASA Sees Fields of Green Spring up in Saudi Arabia”, date of last visit 09/11/2020: https://www.nasa.gov/topics/earth/features/saudi-green.html.

[3] Y. Chen, Y. Wei, Q. Wang, F. Chen, C. Lu, and S. Lei, “Mapping Post-Earthquake Landslide Susceptibility: A U-Net Like Approach”, Remote Sensing, vol. 12, pp. 2767(1-25), 2020.

[4] S. Obata, P. Bettinger, C.J. Cieszewski and R.C. Lowe, “Mapping Forest Disturbances between 1987–2016 Using All Available Time Series Landsat TM/ETM+ Imagery: Developing a Reliable Methodology for Georgia, United States”, Forests, vol. 11, pp. 335(1-16), 2020.

[5] A. Bannari and Z.M. Al-Ali, “Assessing Climate Change Impact on Soil Salinity Dynamics between 1987–2017 in Arid Landscape Using Landsat TM, ETM+ and OLI Data”, Remote Sensing, vol. 12, pp. 2794(1-18), 2020.

[6] X. Cheng, M.M. Nizamani, C.Y. Jim, K. Balfour, L. Da, S. Qureshi, Z. Zhu and H. Wang, “Using SPOT Data and FRAGSTAS to Analyze the Relationship between Plant Diversity and Green of Zhanjiang, China”, Remote Sensing, vol. 12, no. 3, pp. 3477(1-16), 2020.

[7] Z. Yu et al., “Dynamics of Hierarchical Urban Green Space Patches and Implications for Management Policy”. Sensors, Basel, vol. 17, June 2017.

[8] L. Zhang, B. Vermaa, D. Stockwell, and S. Chowdhurya, “Density Weighted Connectivity of Grass Pixels in Image Frames for Biomass Estimation”. Expert Systems with Applications, vol. 101, February 2018.

[9] A. H. Mohammed, A. M. Gadallah, H. A. Hefny, and M. Hazman, “Fuzzy based approach for discovering crops plantation knowledge from huge agro-climatic data respecting climate changes”, Computing, vol. 100, pp. 689-713, April, 2018.

[10] A. A. Dhaif Allah, N. Bin MD. Hashim, and A. Bin Awang, “Discovering Trends of Agricultural Drought in Tihama Plain, Yemen: A Preliminary Assessment”, Indonesian Journal of Geography, vol. 49, pp. 17- 27, June 2017.

[11] A. Choukuljaratsiri, P. Chaovalit, and S. Pongnumkul, “Plant cover detection from visible satellite imagery”. IEEE International Computer Science and Engineering Conference (ICSEC). Thailand, pp. 1-6, November, 2015.

[12] T. J. Hawbaker et al., “Mapping burned areas using dense time-series of Landsat data”, Remote Sensing of Environment, vol. 198, pp 504-522, September 2017.

[13] E. L. Bullock, C. E. Woodcock, and C. E. Holden, "Improved change monitoring using an ensemble of time series algorithms", Remote Sensing of Environment, vol. 238, March 2020.

[14] Z. Zhu, and C. Woodcock, “Continuous Change Detection and Classification of Land Cover Using All Available Landsat Data”, Remote Sensing of Environment, vol. 144, pp. 152-171, March 2014.

[15] G. C. Chow, “Tests of Equality Between Sets of Coefficients in Two Linear Regressions”. Econometrica, vol. 28, pp. 591-605, July 1960.

[16] C. Kwan et al., "Change detection using Landsat and Worldview images", Proc. SPIE. USA, vol. 10986, May 2019.

[17] L. Liu, Z. Jia, J. Yang, and N. K. Kasabov, "SAR Image Change Detection Based on Mathematical Morphology and the K-Means Clustering Algorithm," in IEEE Access, vol. 7, pp. 43970-43978, 2019.

[18] J. Yan et al., “A time-series classification approach based on change detection for rapid land cover mapping,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 158, pp. 249-262, December 2019.

[19] Y. Fu et al., “Characterizing the spatial pattern of annual urban growth by using time series Landsat imagery,” Science of The Total Environment, vol. 666, pp. 274-284, May 2019.

[20] S. Minu and A. Shetty, “A comparative study of image change detection algorithms in MATLAB”, Aquatic Procedia, vol. 4, pp. 1366-1373, 2015.

[21] P. Turaga, R. Chellappa, and A. Veeraraghavan, “Advances in Video-Based Human Activity Analysis: Challenges and Approaches,” Advances in Computers, vol. 80, pp. 237-290, 2010.

[22] B.D. Lucas and T. Kanade, “An Iterative Image Registration Technique with an Application to Stereo Vision”, Proceedings of the DARPA Image Understanding Workshop, Washington, DC, USA, pp. 674–679, April 1981.

[23] B. Horn and B. Schunck, “Determining optical flow”, Artificial Intelligence, vol. 7, pp.185–203, 1981.

[24] M.J. Black and P. Anandan, “The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields”, Computer Vision and Image Understanding, vol. 63, pp. 75–104, 1996.

[25] B. Andres, F.A. Hamprecht, and C.S. Garbe, “Selection of Local Optical Flow Models byMeans of Residual Analysis”, AGM 2007, LNCS 4713, pp. 72–81, 2007.

[26] K.Md. Shahiduzzaman, K.M. Reza, and N. Tazin, “Combined Local Global Optical Flow Approach in Bio-medical Image Sequences for blood flow detection”, International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, pp.468–472, February 2015.

A Decision Support System for Detecting Age and Gender from Twitter Feeds based on a Comparative Experiments

Roobaea Alroobaea[1], Sali Alafif[2], Shomookh Alhomidi[3]
Ahad Aldahass[4], Reem Hamed[5], Rehab Mulla[6], Bedour Alotaibi[7]
Department of Computer Science, College of Computers and Information Technology
Taif University, P. O. Box 11099,Taif 21944, Saudi Arabia[1]
Department of Computer Science, College of Computers and Information Technology, Taif University, Saudi Arabia

Abstract—**Author profiling aims to correlate writing style with author demographics. This paper presents an approach used to build a Decision Support System (DSS) for detecting age and gender from Twitter feeds. The system is implemented based on Deep Learning (DL) algorithms and Machine Learning (ML) algorithms to distinguish between classes of age and gender. The results show that every algorithm has different results of age and gender based on the model architecture and power points of each algorithm. Our decision support system is more accurate in predicting the age and the gender of author profiling from his\her written tweets. It adopts the deep learning model using CNN and LSTM methods. Our results outperform those obtained in the competitive conference s CLEF 2019.**

Keywords—Decision support system; age detection; gender detection; author profiling; deep learning; machine learning

I. Introduction

Recently, many studies have focused on the field of Information Retrieval (IR). The primary target of IR is to extract critical information of a field by predicting the behavior, preference, and person's characteristics. Subsequently, filtering these massive amounts of data to convert it to useful information. This information will be used in the decision support systems. These systems can be used for creating marketing strategies [1], giving viewing suggestions to the user [2], filtering and translate languages [3]. Furthermore, support investigations by studying the sensitive text for national security that detects the source of a threat and helps police in detecting characteristics of the criminal from his\her linguistics [4] [5].

In addition, a decision support system (DSS) plays a great deal nowadays. For bidding websites, there is a needed to know age and gender of the user in order to propose to him\her the service or the clothes. It is becoming a challenging problem. For this reason, the decision support system is a needed by using author profiling.

The aim of this work is to solve the problem of author profiling (AP) by proposing a decision support system (DSS) for detecting age and gender from Twitter Feeds. Author profiling means determining the author's gender and the author's age group of a text by examining his or her writing and identifying the stylistic features. English tweets were taken as input from the PAN-AP-2019 dataset, then generate age and gender based on extraction features distinguish between gender and age class. To achieve the best result, multiple algorithms were implemented such as Deep Learning (like 'Convolutional Neural Network (CNN)', 'Deep Neural Network (DNN)', and 'Long Short-Term Memory (LSTM)' and Machine Learning (like, 'Naïve Bayes (NB)', 'k-Nearest Neighbors (KNN)' 'Decision trees (DT)', 'Support Vector Machines (SVM)', 'Neural Network', and 'Random Forest (RF)'. The structure of the paper is organized as follows; Section 2 presents a brief review on author profile prediction methods. Section 3 introduces our method used to discriminate between authors. Section 4 presents the experiments as well as the evaluations. Finally, the summarization with the conclusion will be mentioned, following by future work.

II. Literature Review

Text classification is depending on statistical procedures and text mining that provide outputs from calculations of extracted terms regularity. Author profiling distinguishes between classes of authors studying their socialist aspect, that is, how people share language. This helps in identifying profiling aspects such as gender, age, native language, or personality type. Author profiling is a problem of increasing significance in applications in forensics, security, and marketing. Table I shows up-to-date researches about detecting age and gender based on DL and ML algorithms.

In addition, there is a need to cover some works that had done by pioneer scientists. The following are some of them:

The author(s) in [14] discovered the possibility of the author automatically classified, which depending on the writer's gender or age using the AP method. AP is predicting features linked to the text author. It includes many dimensions for example gender, age, native language, personality, education level, etc. As stated by [14], men who use more determiners to classify things (this/ this/the, an/an, etc.) and quantifiers (more, little, two, etc.). However, women are more concerned in relations and, thus, using personal pronouns (I, me, her, you, etc.) more than latter. Advanced, [14] modeled the author profiling which is a prediction system. He proposed a technique in his name called the Koppel algorithm. It contains quantifying the duplication of 467 English keywords (such as too, a, too, their, yourself, etc.) in a text to compute the gender of its writer. All writing styles were analyzed from textbooks, fiction, and tests. Then, the program was able to deliver four of five correct answers.

TABLE I. MODERN RESEARCH ON AGE AND GENDER DETECTION

Work	**Method used**	**Publishing Year**	**Text languages**	**Age Result**	**Gender Result**
Alroobaea [6]	NB, Neural Network	2020	English	0.63	0.96
Ashraf and Nie [7]	Bi-LSTM	2020	English Roman-Urdu	0.73	0.88
Rosso and Rangel [8]	RF Logistic regression	2020	Arabic	0.56	0.81
Gishamer [9]	SVM	2019	English	NON	0.89
			Spanish		0.84
Kapočiūtė-Dzikicnė and Damaševičiu [10]	LSTM CNN	2018	Lithuanian	0,316	0.609
Dias and Paraboni [11]	LSTM char n-gram	2018	English	NON	0.66
			Spanish		0.67
			Arabic		0.68
Bayot and Gonçalves [12]	LSTMs	2018	English	NON	0.772
			Spanish		0.687
			Arabic		0.676
Veenhoven et al, [13]	Bi-LSTM CNN	2018	English	NON	79.3
			Spanish		80.4
			Arabic		74.9
Sboev et al [33]	CNN LSTM	2016	Russian	NON	0.86
Mechti et al [3]	DT J48	2013	English	0.58	0.58
			Spanish	0.51	0.556
Gaustad et al [15]	DT J48, RF, IBk, JRip, SMO, Bagging, AdaBoostM1	2007	Arabic	0.7210	0.8115
Tayfun et al [16]	NB Term-based ,k-NN	2006	Spanish	0.51	0.556
	NB Style-based, k-NN				0819
Koppel et al [14]	winnow	2003.	English	NON	0.80

The author in [15] worked on automatic classification of messages in the Arabic language and English languages. They used datasets for three demographic and four psychometric. Authors trait from the linguistic features were trained by the WEKA toolkit and machine learning which includes: 1) decision trees J48, 2) Random Forest, 3) lazy learners IBk, 4) rulebased learners JRip, 5) Support Vector Machines AMO, 6) ensemble/meta-learners Bagging, 7) AdaBoostM1. Bagging looks to income from feature choice, where the SVM based SMO algorithm does not add progress when joint with feature choice. As a result, SMO and Bagging give the best performing for all traits examined. The research got a percentage of 81.5% of well-classified documents about the sex and of 72% about age.

The author in [16] considers the first five problem as a text classification problem. Second, the influence of stylistic features (e.g., word lengths) on predicting the gender. They used a dataset from a chat server (Heaven BBS), when users have peer-to-peer communication via textual messages. The dataset gathering of messages logs storing the users' outgoing messages. They used two different techniques according to gender classes, which are style-based classification and term-based classification. In the Term-Based Classification, they used supervised learning and employed four algorithms (k-NN, naive Bayesian, covering rules, and backpropagation). Every experiment separately was repeated five times. In the Style-Based Classification, they used many stylistic features from the message dataset. The chat messages include emotion-carriers called smileys and emoticons, and they research the word and phrase lengths as a computer-mediated text.

In regarding to the decision support systems (DSS), many systems have been proposed or improved for some areas such as diagnosis of periodontal disease, sales, business ideas competitions, etc. These systems are a computer program used to support determinations, judgments, and ways of action in an organization or a commercial. It collects and analyzes a huge of unstructured data to predict information that can help to get right step in the future or solve problems [36][37]. As mentioned above, the proposed decision support system will help to solve the problem of author profiling. In the next section, our data and methodology will be explained in detailed.

III. CORPUS AND METHODOLOGY

A. Corpus

The data was collected from the CLEF (Cross-Language Evaluation Forum) initiative. It is a self-organized entity whose key goal is to foster science, creativity and the advancement of structures for access to knowledge, with a focus on multilingual and multimodal information at diverse aspects of structure. Since 2010, CLEF has housed evaluation labs for PAN (Plagiarism, Authorship and Near Duplicate Content), offering excellent locations at all times.

Therefore, our corpus is taken from the PAN-AP-2019 dataset. The feeds are taken from Twitter in English language. The dataset is split into two subsets: "training" and "test". The considered dataset is labeled with classes of ages and genders [6]. For age classification, there are four classes (i.e.," 18 -24 "," 25 - 34", "35 - 49" and "50 and more". For gender classification, there are two classes ("Female "and "Male"). There are '14166' tweets that are unbalanced distributed as shown in Table II.

TABLE II. AGE AND GENDER DISTRIBUTION

Class	**Number of Tweets**
Gender Classifications	
Male	6986
Female	7180
Age Classifications	
18-24	5588
25-34	5624
35-49	1865
50-XX	1089

B. Architecture of the System

The proposed system contains the next steps (see Fig. 1).

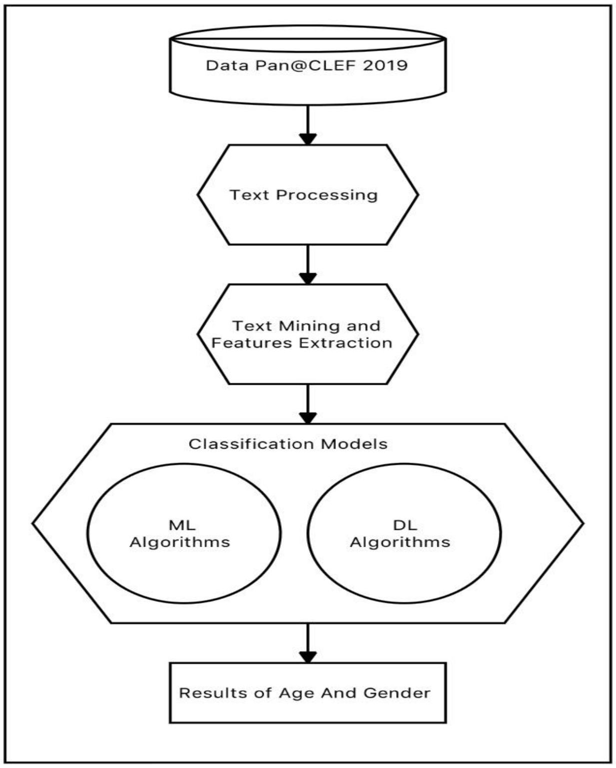


Fig. 1. Construction of the Applied System.

- Text Processing: There are two main problems with the Corpus. First, the model cannot take the tweets directly as input. Secondly, the text could be messy (raw data). As a result, there is a need to converting the raw data into a clean data set by removing noisy data, such as HTML tags, white spaces, inflection words, symbols that would reduce the accuracy of detecting [17].

- Text Mining and Features Extraction: In this step, information that has patterns and trends from the text was extracted by implementing linguistic and statistical processing that study word frequency distributions occurred in the corpus. Then, it sorted these words by frequency [17]. After that gathering these terms into thematic classes to distinguish between age group and gender classes depending on Stylistic features and those based on content [18]. To determine the most word frequency in the corpus and to reduce redundant attributes and to computational capacity to process and to enhance classification performance, TF * IDF were calculated for each class and iterative this step until getting the most discriminating attributes [17], as shown below in Eq.(1) [19]:

$$w_d = f_{wd} \times \log(|D| f_{wd}) \qquad (1)$$

With: d: documents, W: the word, D: one document ∈ D, $f_{w,d}$: number of occurrences of w in d, |D|: the size of the corpora $f_{w,D}$:number of document [17] [19].

- Classification Models: As mentioned above, DL and ML algorithms were used to implement the classification step to gender and age files. Next is explanation for these methods.

1) Deep learning (DL): It is primarily used as a statistical tool for categorizing patterns based on sample data and applying multi-layer neural networks. A group of input units, such as words or pixels, are present in the input layer. The hidden layer includes hidden units (the deeper a network is said to be, the more such layers), and ultimately the output units. With ties running between those nodes. After a while, a back-propagation algorithm allows a process called gradient descent to set the links between units that used a process, so that the same output is generated by any given input [20]. Deep learning methods are used include the following:

- Deep Neural Network (ANN): An artificial neural network technology simulates human brain neural networks activity. It contains an input layer, an output layer, and multiple hidden layers that finds the correct mathematical manipulation to turn the input into the output and flows without looping back [21].

- Convolutional Neural Network (CNN): It is a technology used in image recognition. It consists of fully connected networks that mean each neuron in one layer linking to all neurons in the next layer to analyzing it by using unsupervised learning and adding weights to the loss function [22].

- Long Short-Term Memory (LSTM): It is a class of artificial recurrent neural network (RNN) that address the problem of training RNNs with the technique of back-propagation through time (BPTT). It can process long sequence of dependencies such as speech or video. So, it is considered as the solution to short-term memory in RNN by using distinct units called memory blocks in the recurrent hidden layer [23]. The memory blocks contain [24]:
  - Memory Cells: remembering the historical state of the network. Also, it has weights input, output, and the internal state as following:
    - Input Weight: Weight input for the present time.
    - Output Weight: Weight output from last time.
    - Internal Weight: Internal state applied to calculate the output.
  - Three control gates: in charge of controlling the information flow:
    - Input Gate: Controls the flow of input activations into the memory cell.
    - Output Gate: Controls the output flow of cell activations into the rest of the network.
    - Forget Gate: Decides what information to delete from memory cells.

Input gate and forget gate using to updating of internal state. The output gate is the last filter for output. All these gates with consistent data flow called constant error carrousel (CEC) that help to save cells neither exploding gradient (values of the model’s weights quickly become very large during training) or vanishing gradient.

Most of LSTM advantages are achieved by using back-propagation through time (BPTT) supervised learning algorithm that is used to modify weight to minimize the error of output compared to the expected output in the response of corresponding input [24].

2) Machine learning (ML): It is a research field of artificial intelligence that extract knowledge from data by analyzing them to predict and make the design. The applications of machine learning models become in everyday life from automatic recommendation movies to what food to order or recognizing people in photos and many other applications that contain ML models [25]. The following are ML techniques used in our approach:

- K-Nearest Neighbors (K-NN): It can be defined as a classification algorithm that can predict a new data point by discovering the nearby training set relaying on fixed number 'k' to that point and allocate it to the training set. It is the simplest algorithm to using and understand. It offers best performance without any significant adjustment. Nevertheless, it is lazy and cannot handle several features [25].
- Naïve Bayes Classifiers (NB): They are types of classification algorithm based on the statistical theory of Bayes. They assumed that there is a specific feature in a class that is not related to the presence of any other feature. Their advantages are predicting fast and dealing with large datasets [25].
- Decision Trees (DT): They are types of classification and regression algorithms. They performed classification without requiring much computation. In addition, they provide a clear indication of which fields are most important for prediction or classification. The target of this model is predicting the value of a target variable based on several input variables. They are unlike linear models; they are adaptive to solve any type of problems (classification or regression) [25].
- Random Forest (RF): It is defined as a classification algorithm. It has several decision trees, where each tree is a little different from the others. It solves the problem of over-fitting in training data. The over-fitting can be reduced by being an average of their results. [25].
- Support Vector Machine (SVM): An efficient supervised ML algorithm delivers better performance for both classification and regression exercises. It is accurate and works on a small amount of revised data, and it does not work on a large amount of data because it requires more training time. SVM adapts training data containing features and its preferred class, it contains two stages:
  - Learning Stages: where SVM detects closest data points to the decision boundary known as Support Vectors (SVs), which forms the best separation among the classes [26].
  - Prediction Stages: These SVs are applied to predict the class of test records [26].

C. *Evaluation of the System*

By using the confusion matrix shown in Table III, precision, accuracy, recall, f-score can be calculated with the equations (2), (3),(4), and (5) that evaluate the performance for our unbalanced dataset on ML algorithms [27]. The accuracy for DL algorithms were calculated [28].

TABLE III. CONFUSION MATRIX

		Predicted class	
True Class	**Class A**	**Class A**	**Class B**
		True Positives (TP)	False Negatives (FN)
	Class B	False Positives (FP)	True Negatives (TN)

Where TP (True Positives) and TN (True Negatives) positive and negative correct labeled predictions. In Opposition, False Positives (FP) and False Negatives (FN) are positive and negative incorrect labeled predictions.

- Accuracy: Measure how the system is correctly predicted [34].

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

- Precision: The number of true positives divided by all (positive) predictions [34].

$$Precision\ = \frac{TP}{(TP+FP)} \quad (3)$$

- Recall: Percentage of positive predictions in the system that correctly identified [34].

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

- F_score: Measure the harmonic between recall and precision [34].

$$F\_score\ = 2 \times \frac{Precision \times Recall}{precision+Recall} \quad (5)$$

IV. Experimentation and Results

In this research, DL and ML algorithms were implemented to detect the age and gender of an author. DL algorithms results are more efficient for our model rather than ML algorithms. In the first time of implementing our approach, we did not get as expected results from CNN age's code, also LSTM age and gender codes.

Notice that using either "sigmoid" or "softmax" as an activation function will not affect gender because we have only two classes called "Binary Classification" [31]. On the other hand, we used "softmax" for age's code "Multi Class Classification", but we had the same accuracy and loss in each step in both models. According to this problem, the solution is mentioned at [32]. We had changed the activation function to "sigmoid", but this provides loss results in negative. We decide to change the loss function from "binary_crossentropy" to "mean_squared_error". This provides good accuracy as shown below in (Table IV). As a result, the searched was done for the reasons that causing these problems and solved them as shown below:

A. LSTM Gender's Code Problem and Solution

By using 50000 words as top frequent words in LSTM. In addition to 10 hidden cells in the "Dense" layer. Had not provided good accuracy as in CNN algorithm. As a result, the number of hidden cells was increase to 20 "Dense" layer and number of frequent words was decreased into 5000 as shown in this problem [29]. Our approach had increased the accuracy from 0.5068 to 0.9906 as shown below in (Table V).

B. Age's Code Problem and Solution

Convolutional Neural Network and Long Short-Term Memory algorithms had provided great results for gender code according to the balance of gender tweets as shown in Table VI. Table IV shows how we implemented "binary_crossentropy" loss function to the model, that is used when there are only two label classes (Male and Female) as this was mentioned at [30]. In addition to using "sigmoid" as "activation" function.

Several classification models were tried as shown in (Table IV) and (Table V); the best result is achieved by implementing Convolutional Neural Network CNN as the first place and LSTM algorithms as the second place, in both age and gender detection. According to our search, we believe that if our model had a larger dataset size; the LSTM algorithm will be more powerful rather than the CNN algorithm, as mentioned in [35]. In Table VI, our approach is compared with the best accuracy result achieved of gender detection of competition "PAN at CLEF 2019" implemented by Valencia et al team for English tweets. It is obvious that our model system obtains more accurate results than others in predicting gender from author profiling.

TABLE IV. Age Detection Results

Deep Learning Algorithms					
Algorithm		Epoch Numbers	Accuracy	Loss	
CNN		30	0.7489	0.5210	
LSTM		30	0.7265	0.5434	
Machine Learning Algorithms					
Algorithm		Precision	Recall	Accuracy	F-score
K-NN	0	0.61	0.61	0.50	0.61
	1	0.66	0.40		0.50
	2	0.38	0.50		0.43
	3	0.16	0.39		0.23
NB	0	0.61	0.89	0.62	0.72
	1	0.63	0.67		0.65
	2	1.00	0.04		0.07
	3	0.00	0.00		0.00
DT	0	0.62	0.68	0.58	0.65
	1	0.59	0.61		0.60
	2	0.49	0.38		0.43
	3	0.34	0.25		0.29
SVM	0	0.66	0.83	0.66	0.74
	1	0.64	0.71		0.68
	2	0.82	0.32		0.47
	3	0.84	0.17		0.28
Neural Network	0	0.69	0.69	0.63	0.69
	1	0.63	0.72		0.67
	2	0.54	0.41		0.46
	3	0.35	0.24		0.28
RF	0	0.63	0.88	0.64	0.73
	1	0.63	0.65		0.64
	2	0.94	0.26		0.41
	3	1.00	0.06		0.11

TABLE V. GENDER DETECTION RESULTS

<table>
<tr><td colspan="6"><b>Deep Learning Algorithms</b></td></tr>
<tr><td><b>Algorithm</b></td><td colspan="2"><b>Epoch Numbers</b></td><td><b>Accuracy</b></td><td colspan="2"><b>Loss</b></td></tr>
<tr><td>CNN</td><td colspan="2">30</td><td>0.9965</td><td colspan="2">0.0023</td></tr>
<tr><td>LSTM</td><td colspan="2">30</td><td>0.9906</td><td colspan="2">0.0211</td></tr>
<tr><td colspan="6"><b>Machine Learning Algorithms</b></td></tr>
<tr><td rowspan="2">K-NN</td><td>0</td><td>0.59</td><td>0.66</td><td rowspan="2">0.61</td><td>0.62</td></tr>
<tr><td>1</td><td>0.63</td><td>0.56</td><td>0.59</td></tr>
<tr><td rowspan="2">NB</td><td>0</td><td>0.75</td><td>0.56</td><td rowspan="2">0.69</td><td>0.64</td></tr>
<tr><td>1</td><td>0.66</td><td>0.82</td><td>0.73</td></tr>
<tr><td rowspan="2">DT</td><td>0</td><td>0.61</td><td>0.59</td><td rowspan="2">0.61</td><td>0.59</td></tr>
<tr><td>1</td><td>0.61</td><td>0.63</td><td>0.62</td></tr>
<tr><td rowspan="2">SVM</td><td>0</td><td>0.70</td><td>0.66</td><td rowspan="2">0.69</td><td>0.68</td></tr>
<tr><td>1</td><td>0.69</td><td>0.72</td><td>0.70</td></tr>
<tr><td rowspan="2">Neural Network</td><td>0</td><td>0.69</td><td>0.68</td><td rowspan="2">0.69</td><td>0.69</td></tr>
<tr><td>1</td><td>0.69</td><td>0.70</td><td>0.70</td></tr>
<tr><td rowspan="2">RF</td><td>0</td><td>0.71</td><td>0.62</td><td rowspan="2">0.69</td><td>0.66</td></tr>
<tr><td>1</td><td>0.67</td><td>0.75</td><td>0.71</td></tr>
</table>

TABLE VI. COMPARED BEST RESULTS

Teams	**Method Used**	**Best Gender Accuracy Result**
Our Team	CNN	0.9965
	LSTM	0.9906
Valencia et al [36]	SVM Logistic Regression Multi-Layer Perceptron	0.8432

V. CONCLUSION

In this research, DL and ML algorithms were used to solve the PA problem from English tweets that was taken from the PAN-AP-2019 dataset. To achieve the best result, multiple algorithms of Deep Learning were implemented, which are DNN, CNN and LSTM. In addition, Machine Learning algorithms were implemented too, which are "KNN, NB, DT, SVM, Neural Network, and RF". The results showed that every algorithm has different results of age and gender based on the model architecture and power points of each algorithm.

Moreover, our decision support system and model are achieved good result than others, who participated in the competitions CLEF 2019 (Table VI). It is based on deep learning model using CNN and LSTM methods. It is more accurate in predicting the age and the gender of author profiling from his\her written tweets. Then, it will be great if it is adopted in any bidding websites that need to know the age and the gender of their users in order to propose to them the service such as book or clothes, etc.

REFERENCES

[1] E. Lundeqvist,and M. Svensson. Author profiling: A machine learning approach towards detecting gender, age and native language of users in social media, 2017.

[2] PranavDar/SEPTEMBER11,2018/analyticsvidhya/february,29,2020/https://www.a nalyticsvidhya.com/blog/2018/09/facebook-rosetta-process-text-billions-images/. Acecc [22/2/2020].

[3] S. Mechti,, M. Jaoua , L. H. Belguith, & R, Faiz.. Author profiling using style-based features. Notebook Papers of CLEF2, 2013.

[4] S.S. Yatam, T.R. Reddy. Author profiling: Predicting gender and age from blogs, reviews & social media. International Journal of Engineering Research & Technology (IJERT), ISSN, pp.2278-0181. 2014.

[5] Tom Buchanan, A. John Johnson, and R. Lewis Goldberg. Implementing a Five-Factor personality inventory for use on the internet. European Journal of Psychological Assessment, 21:115–127, 2005.

[6] R. Alroobaea. An Empirical combination of Machine Learning models to Enhance author profiling performance. International Journal, 9(2), 2020.

[7] M.A. Ashraf, R.M.A Nawab, and F. Nie, Author profiling on bi-lingual tweets. Journal of Intelligent & Fuzzy Systems, (Preprint), pp.1-11.2019.

[8] P. Rosso, F. Rangel. Author Profiling Tracks at FIRE. SN Computer Science, 1(2), pp.1-11, 2020.

[9] F, Gishamer. Using Hashtags and POS-Tags for Author Profiling Notebook for PAN at CLEF , 2019.

[10] J.Kapočiūtė-Dzikicnė and, R. Damaševičius. Lithuanian author profiling with the deep learning. In 2018 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 169-172). IEEE. 2018.

[11] R.F.S. Dias, and I. Paraboni. Author Profiling using Word Embeddings with Subword Information: Notebook for PAN at CLEF 2018. In CLEF (Working Notes), 2018.

[12] R.K. Bayot, and T. Gonçalves. Multilingual author profiling using lstms. In Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018) , 2018.

[13] R. Veenhoven, S. Snijders, D. van der Hall and R van Noord. Using translated data to improve deep learning author profiling models. In Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018) (Vol. 2125), 2018.

[14] M. S. Koppel and A. Shimoni, Automatically categorizing written texts by author gender, Literary and Linguistic Computing, pages 401-412, 2003.

[15] T. Gaustad, D. Estival and B. Hutchinson. TAT: an author-profiling tool with application to Arabic emails. Proceedings of the Australasian Language Technology Workshop, pages 21-30, Melbourne, Australia, 2007.

[16] B. Tayfun Kucukyilmaz, B. Barla Cambazoglu, Cevdet Aykanat, and Fazli Can. "Chat mining for gender prediction." In International conference on advances in information systems, pp. 274-283. Springer, Berlin, Heidelberg, 2006.

[17] R. Alroobaea, A.H. Almulihi, F. S. Alharithi, S., Mechti, M .Krichen. and L.H. Belguith,, A Deep learning Model to predict gender, age and occupation of the celebrities based on tweets followers. 2020.

[18] J. Pennebaker . The secret life of pronouns: What our words say about us. New York,USA, 2011.

[19] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval, information processing and management, pages 513--523, 1988.

[20] G. Marcus. Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00631. 2018.

[21] M.J. Willis, C. Di Massimo, G.A. Montague, M.T. Tham and A.J. Morris. May. Artificial neural networks in process engineering. In IEE Proceedings D (Control Theory and Applications) (Vol. 138, No. 3, pp. 256-266). IET Digital Library. 1991.

[22] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun and R. Fergus. Regularization of neural networks using dropconnect. In International conference on machine learning (pp. 1058-1066). 2013.

[23] H. Sak, A. Senior and F. Beaufays. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. arXiv preprint arXiv:1402.1128. , 2014

[24] J. Brownlee. Long Short-term Memory Networks with Python: Develop Sequence Prediction Models with Deep Learning. Machine Learning Mastery , 2017.

[25] A.C. Müller and S. Guido. Introduction to machine learning with Python: a guide for data scientists. "O'Reilly Media, Inc.".2016.

[26] A. Kulkarni, Y. Pino, & T. Mohsenin,. SVM-based real-time hardware Trojan detection for many-core platform. In 2016 17th International Symposium on Quality Electronic Design (ISQED) (pp. 362-367). IEEE. 2016.

[27] A. Alsufyani, R. Alroobaea & K.A. Ahmed. Detection of single-trial EEG of the neural correlates of familiar faces recognition using machine-learning algorithms. International Journal of Advanced Trends in Computer Science and Engineering. [Online] 8 (6). Available from: doi:10.30534/ijatcse/2019/28862019. 2019.

[28] M. Aykanat,and P. Bendevis. Keras Accuracy Does Not Change. [online] Stack Overflow. Available at: <https://stackoverflow.com/ questions/37213388/keras-accuracy-does-not-change> [Accessed 4 November 2020].

[29] TensorFlow. Tf.Keras.Losses.Binarycrossentropy | Tensorflow Core V2.3.0. [online] Available at: <https://www.tensorflow.org/api_docs/ python/tf/keras/losses/BinaryCrossentropy>.

[30] b. chandra, P. Nayak and n. swami. How Does Sigmoid Activation Work In Multi-Class Classification Problems. [online] Data Science Stack Exchange. Available at: <https://datascience.stackexchange .com/a/ 72731> [Accessed 4 November 2020].

[31] M. Aykanat, and P. Bendevis. Keras Accuracy Does Not Change. [online] Stack Overflow. Available at: <https://stackoverflow.com/ questions/37213388/keras-accuracy-does-not-change> [Accessed 4 November 2020].

[32] A., Sboev, T., Litvinova, D., Gudovskikh, R., Rybka, & I. Moloshnikov. Machine learning models of text categorization by author gender using topic-independent features. Procedia Computer Science, 101, 135-142, 2016.

[33] W. Ahmed and M., Bahador. The accuracy of the lstm model for predicting the s&p 500 index and the difference between prediction and backtesting. 2018.

[34] S. Mechti, M. Jaoua, R. Faiz, H. Bouhamed, L.H. Belguith. Author Profiling: Age Prediction Based on Advanced BayesianNetworks. Research in Computing Science; 110: 129–137.2016

[35] A.I.V. Valencia, H.G. Adorno, C.S. Rhodes, and G.F. Pineda. Bots and Gender Identification Based on Stylometry of Tweet Minimal Structure and n-grams Model. 2019.

[36] M., Farhadian , P. Shokouhi and P. Torkzaban. A decision support system based on support vector machine for diagnosis of periodontal disease. BMC Research Notes, 13(1), pp.1-6. 2020.

[37] D., Martins, R., Assis, R. Coelho and F. Almeida. Decision Support System for Business Ideas Competitions. Journal of Information Systems Engineering and Management,4(3), 2019.

Admission Exam Web Application Prototype for Blind People at the University of Sciences and Humanities

Alexis Carrion-Silva[1], Carlos Diaz-Nuñez[2], Laberiano Andrade-Arenas[3]
Faculty of Science and Engineering
Universidad de Ciencias y Humanidades, Lima Perú

***Abstract*—Currently, there is a large sector of Peru's population that has some type of disability. Every year, the government creates norms for their integration into society. However, to date, total integration is not achieved. One of the points that can be seen is at the moment of taking an entrance exam, where to this day, they do not have the tools to perform in an autonomous and optimal way. Taking this into account, the objective of this article is the development of a prototype admission test for blind people at the University of Sciences and Humanities. A hybrid methodology is used with between Soft Systems and Scrum. The results were obtained from the analysis of both methodologies and a final product prototyped with Balsamiq, demonstrating an optimal union between these two methodologies. Therefore, the prototype will facilitate the performance of blind people during their entrance exam.**

Keywords—Admission; Balsamiq; blind; scrum; soft systems

I. Introduction

The entrance exam is one of the fundamental requirements for the entrance of students to universities today, although it is true that some universities have more rigorous entrance exams than others, specifically public universities. Today, many of these universities do not have admission test modalities for the visually impaired or tools that allow them to perform optimally when taking the admission test.

There are currently 39 million visually impaired people worldwide, representing 13% of the world's population [1]. Consequently, the technological evolution aimed at people with visual impairment has grown considerably, largely due to their difficulties. Today there are a great number of technological resources that guide people with visual impairment in their tasks, improving their autonomy in various aspects.

Online dating applications, is an example that nowadays has become a very common way to meet people. However, in most online dating platforms, most of the data is visual [2]. This creates an access barrier for people with visual impairments and prevents them from being able to fully use this tool or any type of online platforms.

In Peru, the National Institute of Statistics and Information (INEI) reported that 801,000 people are permanently limited in their ability to see, even when using glasses. Moreover, 52.6% are in urban areas and 44.8% in rural areas [3]. However, this number has not yet been fully included in many aspects of society, especially in university education, since tests for them are not yet well implemented. For example, the representatives in Puno noticed the absence of staff in charge of reading the questions to blind applicants [4]. He made it clear that even if they take the exams physically, they are still affected and do not have a better chance of taking the exam.

The inclusion in the applications should be guided in a participatory and collaborative act. Human action must be guided by dialogue between differences; the perspective of social-digital inclusion demands a change in the technological profile that must be a goal [5]. Under this concept of inclusion, the creation of a prototype application was thought of, which would make it possible to answer an online test in a simpler way for a blind person. This will allow a great step towards inclusion and will help to have greater chances of success in it.

The objective of this article is to develop a web prototype so that blind applicants can take an admission exam at the University of Sciences and Humanities with greater autonomy. In this way, a great step towards inclusion for the visually impaired will be taken.

The article is divided into sections, where Section II details the methodology to be implemented; Section III the application of the case study; Section IV the results and discussions obtained and Section V the conclusions and future work.

II. Methodology

For the elaboration of the web prototype, a hybrid methodology will be used, implementing stages 1 and 2 of the soft systems methodologies in order to identify in depth the problem to be addressed. Similarly, the scrum methodology was applied in order to manage the development of the project in an agile way.

A. *Soft Systems Methodology*

This methodology is based on the systemic thinking, where it allows to see under a holistic look, the integration of all the parts as a whole; the first and second stage of this methodology were taken into account: the unstructured and structured situation respectively. In this way, it is possible to analyze the behavior of the problem under study [6].

B. *Scrum Methodology*

The Scrum methodology is a framework for agile development, which facilitates the control and management of software development. It is an iterative and incremental model for the development of software [7]. As we can see in Fig. 1, it shows an example of the Scrum work cycle.

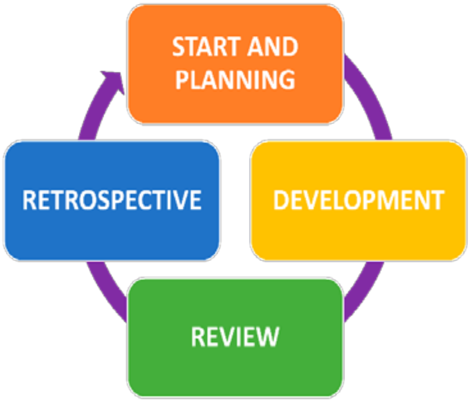


Fig. 1. Scrum Work Cycle.

C. *Scrum Methodology*

This hybrid methodology consists of 6 steps: unstructured stage, structured stage, initiation and planning, development, review and retrospective of the sprint.

1) Unstructured stage: The main causes found in the research problem are described, the central problem being the application of university entrance exams.

2) Structured stage: In this part, all those involved in the problem are described with graphics, interrelating each of them with the application of admission tests to blind people.

3) Initiation and planning of the sprint: The elaboration of the sprints can be considered as mini projects within the whole project. Within the first meeting, the team defines all aspects of its functionality, its objectives, risks, delivery times, among others. Later, meetings are scheduled between the Product Owner and the development team to detail how each point of the sprint will be developed. Changes are evaluated; decisions are made and improvements are made within the sprint [8].

4) Development of sprint: When the sprint is underway, the development team must ensure that the deadlines set in the development of the sprint are met. Similarly, the scope may be renegotiated between the development team and the product owner [8].

5) Sprint review: This is done at the end of each Sprint to check progress and adjust if necessary. Within the review, topics such as: verification of the software by owner, estimated completion dates according to the progress, a timeline review, among others, are seen [7].

6) Retrospective of the sprint: Here, the whole team makes an inspection of itself and a plan of improvements is elaborated to be able to execute them in the next Sprint. They review how they did during the sprint, identify and sort out what went well or badly, and implement the improvement plan to improve the team [7].

D. *Development Tools*

This section presents the tools that were used for the development of the web, which are the Apache server, the MySQL database manager and the PHP programming language.

1) Sublime text: Sublime Text, being a text and source code editor, will allow us to develop the web application, in the language of PHP and the revision of queries.

2) PHP: PHP is an open source programming language natively focused on web development; its application goes hand in hand with HTML. This programming language has a syntax similar to the C language, java and Perl being one of the easiest programming languages to learn [9].

3) Apache: Apache HTTP Server is a free, open source web server software for Unix platforms that runs most websites worldwide. It is maintained and developed by the Apache Software Foundation.

4) MySQL: MySQL is a server and database manager; its architecture makes it a multiplatform tool. MSQL also offers flexibility when working in demanding environments such as web applications. In turn, MySQL can process data in embedded applications, data storage among others, as well as in redundant systems of high availability of online transaction processing (OLTP) [10].

E. *Tools for Prototyping*

Balsamiq and Adobe XD will be used for the elaboration of the application prototypes.

1) Adobe XD: This tool will design the screens or GUI of the web application. Adobe XD is a tool that provides what is needed to design and prototype websites, mobile applications, voice interactions, touch screens and other types of user interactions [11].

2) Balsamiq: Balsamiq Mockups, is a tool for the creation of mockups, it facilitates the creation of the necessary design sketches in the software development process characterized by the tools it has, UI elements library, work area, navigation panel among others [12].

III. Case Study

In this section, the six steps of the hybrid methodology for the problem will be developed.

A. *Unstructured Stage*

This section describes the main causes found in the research problem.

The first point is that there is little access to technology for blind people because, despite the fact that in Peru there are more than 1.5 million people who are blind or visually impaired, there are still not enough tools for them to use [13]. This can be seen in the admission exams, where not all universities have the tools to allow blind people to have a certain amount of autonomy when taking the exam.

The second point is the education law, which, despite establishing that it is a fundamental right of society (Article 3) and that the right to an inclusive education is recognized for all persons with a disability without discrimination (Article 21) [14], in addition, the process of admission to universities in Peru includes persons with a disability, as mentioned in the New University Law [14]. In addition, the admission process to universities in Peru includes persons with disabilities, as mentioned in the New University Law No. 30220 in Article 98: "Persons with disabilities are entitled to a reservation of 5 percent of the vacancies offered in their admission procedures (sic) [15]. In the same way, it was detected that the tests for blind applicants did not arrive on time to the classrooms where the exam was given.

The third point is the inequality that exists today in Peru, not only in education but in many areas of society. Despite the years the state has not been able to close this gap that would allow it to open the way to the inclusion of all citizens, despite its efforts and laws that were created so that this inequality does not exist to date, there are disabling barriers that still exist in society despite the laws and agreements to make a more inclusive world where everyone can feel welcome and accepted [16].

B. Structured Stage

As can be seen in Fig. 2, all those involved in the application of the entrance examinations can be seen; in addition, the interaction between the main parties involved, interpersonal relations and community participation can be seen [17].

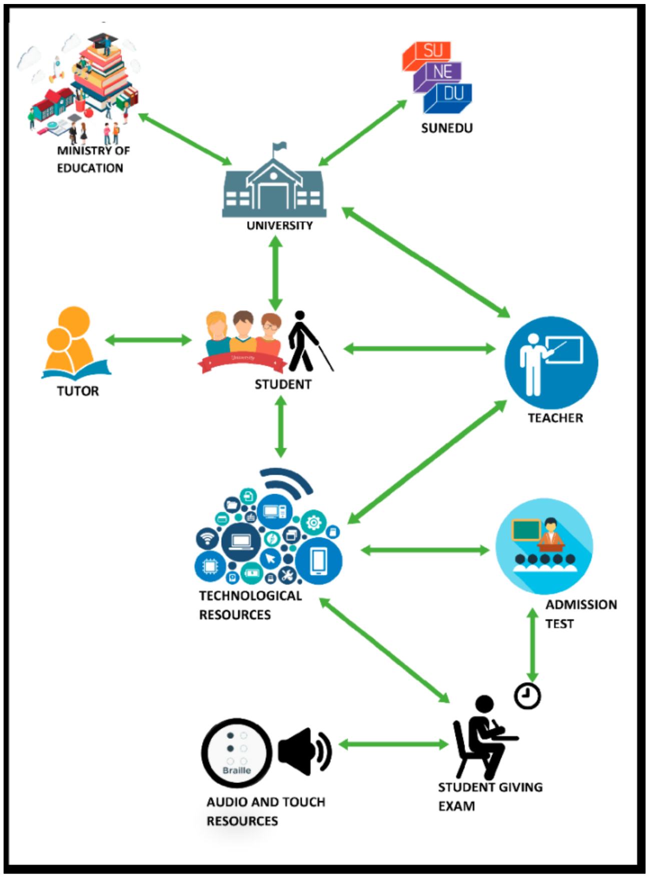


Fig. 2. Structured Situation.

C. Planning and Estimation of user Stories

Table I shows the prioritized user stories, previously estimated using the Delphi estimation group technique; the key words for the stories are "like", "want" and "for".

Having the user stories, we estimated and established Table II of the duration of each Sprint.

D. Development of Sprint and Prototypes

At this stage, the development of the prototypes was carried out according to the modules already established, with Fig. 3, Fig. 4, Fig. 5, Fig. 6, Fig. 7 and Fig. 8.

For each Sprint, five increments were developed, which are the sum of all the elements developed during the Sprint:

Increment 1 (Login Module): Here is developed the interface that allows access to the platform, entering the respective credentials, as shown in Fig. 3.

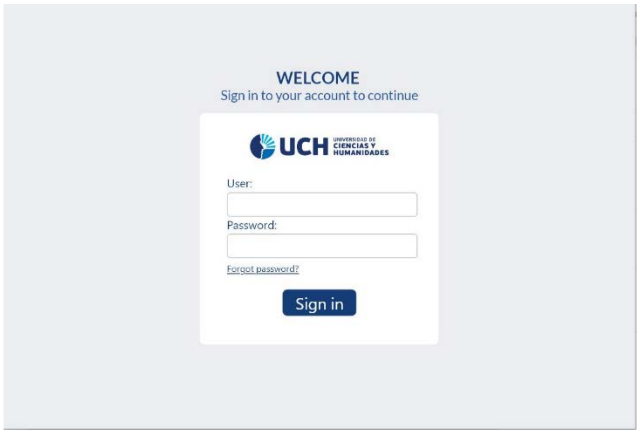
Increment 2 (Registration Module): Fig. 4 shows us this module, which allows us to register applicants so they can take the admission test.

TABLE I. BACKLOG CREATION

User History	Priority
As a user I want to enter the credentials to have access to the admission test platform.	3
As an academic assistant I want to consult, update and register applicants so that they can take their entrance exam.	1
As an academic assistant I want to record the questions and answers of the exam so that applicants can take the exam.	1
As an applicant I want to take the entrance exam in order to get into the university.	1
As an academic assistant I want to generate the results of the entrance exam to publish the entrants.	2
As an academic assistant I want to generate the results of the entrance exam to publish the entrants.	2
As a teacher I want to read the questions on the entrance exam to guide the student in case he or she does not understand the question.	3
As an academic assistant I want to see in graphic format the results of the exam in order to make an analysis of the performance of the applicants.	3
As an academic assistant I want to visualize if any of the applicants have any limitations or disabilities so that I can assign them to a different type of exam.	3

TABLE II. BACKLOG CREATION

User History	Priority
Admission test page for blind people.	14 weeks
Access Module.	2 weeks
Registration Module.	2 weeks
Examination Preparation Module.	4 weeks
Admission Test Module.	4 weeks
Results Module.	2 weeks



Sign in to your account to continue

Fig. 3. Login Screen.

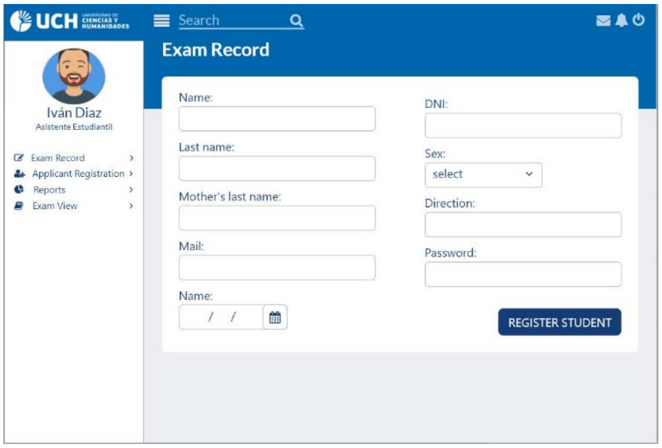


Fig. 4. The Student Registration Screen.

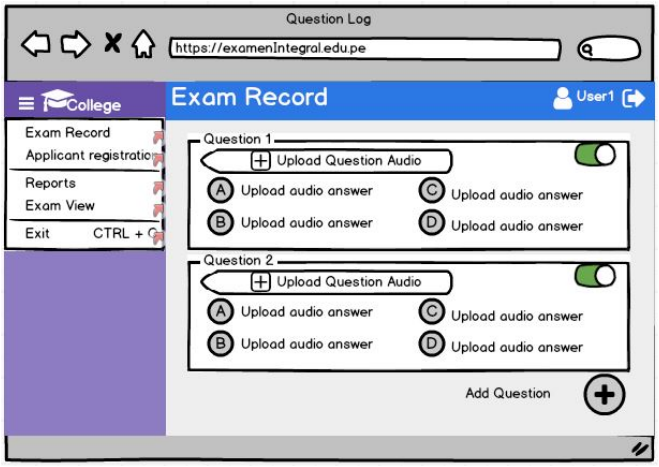
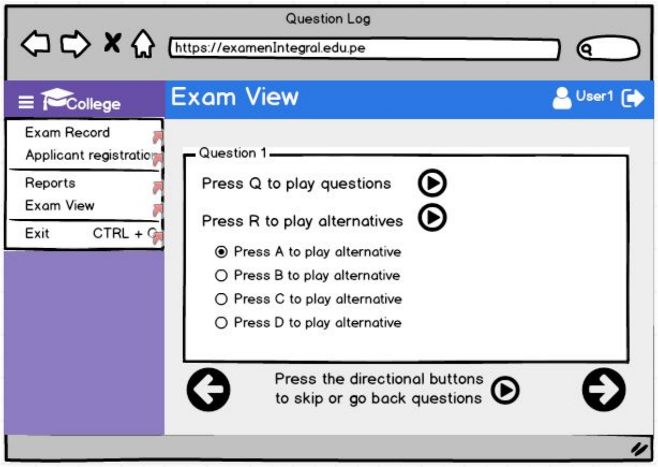


Fig. 5. The Question Registration Screen.

Increment 3 (Admission Test Preparation Module): This increment allows us to record the audios of the admission test questions and answers, as shown in Fig. 5.

Increment 4 (Admission Test Module): In this increment the student can play the audios of the admission test questions and answers, as shown in Fig. 6.

Increment 5 Increment 5 (Results Module): The module allows us to see two options: in the first one the applicant can know what his results were, as shown in Fig. 7; in the second one, the university staff can elaborate different reports on the results of all the applicants, as exemplified in Fig. 8.



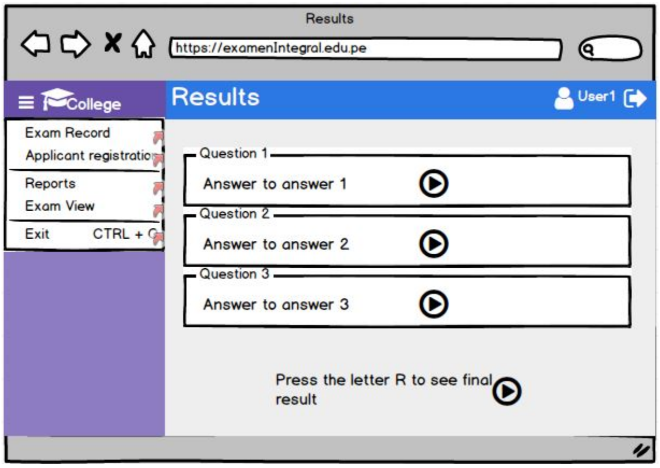
Press Q to play questions

Press R to play alternatives

● Press A to play alternative
○ Press B to play alternative
○ Press C to play alternative
○ Press D to play alternative

Press the directional buttons to skip or go back questions

Fig. 6. The Admission Test Screen.



Press the letter R to see final result

Fig. 7. Admission Test Results Screen.

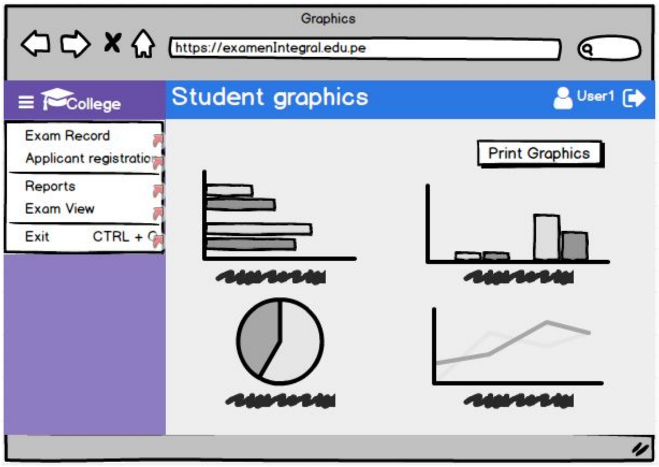


Fig. 8. Admission Test Graphics Screen.

E. Review of the Sprint

The development team is responsible for reviewing the Sprint when it is finished, with an estimated time of three hours maximum. One of the team evaluates that each user history is fulfilled and applies changes if necessary. The team explains the development of each module and mentions each problem that arose during the development and the solution that was applied.

F. Retrospective of the Sprint

For the retrospective, the scrum team evaluates the finished sprint and the techniques that were used during its development. Both the programming language PHP, the database manager MySQL and the prototyping tools Balsamiq and Adobe are evaluated. Members suggest new methods and solutions for the improvement of the next Sprint.

IV. Results and Discussions

A. About the Case Study

The design and development of this case study resulted in the creation of the prototype of the web application for the admission exam for blind people, which has the purpose and objective of showing the correct function, interaction and comfort of the users, ensuring its operation and functionality. A result is presented to provide a solution to the lack of software tools for people with disabilities, thus achieving a step towards integration into society, thus improving the quality of life of this population.

The Balsamiq tool was of great importance, since with it the prototypes of the web application were made, being easy to use and to adapt for the designer; demonstrating its great similarity to a web page. By having some basic options on buttons, grouping and positioning of various components, it allows agility and changes quickly [18]. In the same way, it allowed an easy adaptation when making designs in Adobe XD.

In a comparison with other articles with reference to the applications for admission tests, it was observed that they offer adequate learning environments, tests and feedback, improving the knowledge and the self-evaluation process [19], being a benefit for the student; however, applications like these are not designed or destined to people with visual limitations, because of the visual tools that are used and not sound tools like recordings or texture tools like the friar, being these characteristics clear differences between our prototype and other similar applications. And this can be observed in the record of questions and alternatives Fig. 5, in which the questions and alternatives are entered in audios for people who are heard by the applicants with visual limitations or blind people. The result is a test in which the questions and alternatives can be heard without the need for someone to read them to them, improving their performance at the time of taking the test; the model of the test is shown in Fig. 6. It will consist of audios that will be played at the time of passing the questions, which will be played automatically, indicating the question and the alternatives, as well as the buttons to be pressed to select the alternative. It is worth mentioning that the keyboard will have Braille stickers so that the student can distinguish the letters on the keyboard.

B. About the Methodology

Making a comparison with other articles, the development of the application was made taking into account points of the soft systems methodology, which is shown as an approach to address problematic and disordered situations of all kinds focusing the development [20]; the soft systems methodology compared to other methodologies, such as Design Thinking, has the advantage of analyzing with a holistic vision under the systemic thinking, which allowed articulating all the involved of the problem; on the other hand, Design Thinking does not articulate all the parts. In the same way, for the development of the case study, the steps specified in the Scrum methodology were followed, focusing more on the deliverables. Because having a single model as a waterfall or prototype for development is not enough for the requirements of the product and, therefore, agile development is more useful for the development of customized products [21].

The use of the Scrum methodology allowed the development team and the end users to work with a better productivity, through continuous revisions during each Sprint and with the feedback that helps to solve problems at the time of development. In addition, he demonstrated many advantages of using this development methodology, is that it largely drives collaborative work, accept the changes within each Sprint completed and make software deliveries gradually, until the final product. Similarly, the flexibility it had for the selection of sprints, makes it possible to work in different ways development projects [22].

V. Conclusions and Future Work

The prototype web application helps blind people to take an entrance exam with more autonomy. The application will allow the inclusion of visually impaired people towards an admission test, making it easy to get into any college. The use of the programming language PHP and the database manager MySQL, being free software, reduces the investment cost for the development of the web application. It is recommended that in the future the web application be implemented in Peruvian universities, since it would be a great step towards inclusion for blind people. Similarly, not only use it in the admission exam, but also in the exams they may take throughout their career.

References

[1] D. Rocha, V. Carvalho, E. Oliveira, J. Gonc ̧alves, and F. Azevedo, “Myeyes-automatic combination system ofclothing parts to blind people: First insights,” pp. 1–5, 2017.

[2] M. Zhou, W. Li, and B. Zhou, “An iot system designfor blind,” pp. 90–92, 2017.

[3] INEI, “En el Perú 1 millón 575 mil personas presentan algún tipo de discapacidad,” Extraído de: https://www.inei.gob.pe/prensa/noticias/en-el-peru-1-millon-575-mil-personas-presentan-alg/, 2013.

[4] Defensoria del Pueblo, “Deficiencias en atención a postulantes con discapacidad en examen de admisióna una puno,” 2019.

[5] R. P. Machado, D. Conforto, and L. Santarosa, “Perception for cooperation: Case study in web text editorsfrom the perspective of blind users,” pp. 1–6, 2016.

[6] N. A. Miftahul Huda and I. Sembiring, “The use of softsystems methodology to resolve hoax news problems indonesia,” pp. 65–68, 2018.

[7] F. Hayat, A. U. Rehman, K. S. Arif, K. Wahab, and M. Abbas, “The influence of agile methodology (scrum) on software project management,” pp. 145–149, 2019.

[8] B. L. Romano and A. Delgado Da Silva, “Projectmanagement using the scrum agile Method: A case study within a small enterprise,” pp. 774–776, 2015.

[9] HEURTEL, Olivier. PHP 5.6: desarrollar un sitio web dinámico e interactivo. Ediciones ENI, 2015.

[10] J. S. Aleman Fierro, “Diseñoo de una metodología para cluster de base de datos oracle mysql de alta disponibilidad, con un demo de aplicación en servidores linux.”, B. S. thesis, Quito: UCE, 2016.

[11] C. L. Loor, J. O. Estrada, N. Q. Sanmartin, and G. M.Guacho, “Prototipo de una aplicación móvil para el diseño de curva de carreteras,” vol. 3,no.1, pp. 836–847, 2019.

[12] M. R. Z. NAVA et al., “Análisis de herramientas para el diseño de mockups.,” 2017.

[13] A. Aldaz and P. Juan, “Sistema electrónico para la enseñanza del lenguaje braille a personas invidentes,”2016.

[14] M. V. M. Carrasco, “La discapacidad en el Perú y adaptaciones de accesibilidad de espacios e infraestructura en centros educativos inclusivos,” Educación, vol. 24, no. 1, pp. 35–45, 2018.

[15] M. D. l. A. Terrazas Garcia, “Accesibilidad a la información de usuarios con discapacidad visual a la biblioteca central “pedro zulen” de la unmsm,” 2018.

[16] V. A. Corzo, “Reflexión crítica de la educación inclusiva,” 2020.

[17] S. C. Simplican, G. Leader, J. Kosciulek, and M. Leahy, “Defining social inclusion of people with intellectualand developmental disabilities: An ecological model ofsocial networks and community participation,” Researchin developmental disabilities, vol. 38, pp. 18–29, 2015.

[18] A. Delgado and J. Sosa, “Mobile application design ofgeolocation to collect solid waste: A case study in lima,peru,” pp. 1–4, 2019.

[19] A. Robu, I. Filip, R. Robu, I. Szeidert and C. Vasar, "Online Platform for University Admission," 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA), Zakynthos, Greece, 2018, pp. 1-5, doi: 10.1109/IISA.2018.8633616.

[20] P. Checkland and J. Poulter, “Soft systems methodology,” in Systems Approaches to Making Change: APractical Guide, Springer, 2020, pp. 201–253.

[21] A. Srivastava, S. Bhardwaj, and S. Saraswat, “Scrummodel for agile methodology,” in 2017 International Conference on Computing, Communication and Automation (ICCCA), IEEE, 2017, pp. 864–869.

[22] A. Srivastava, S. Bhardwaj, and S. Saraswat“ Scrummodel for agile methodology,” pp. 864–869, May 2017.DOI: 10.1109/CCAA.2017.8229928

Credit Card Business in Malaysia: A Data Analytics Approach

Mohamed Khaled Yaseen[1]
MSc Data Science, School of Computing
Asia Pacific University of Technology & Innovation
Kuala Lumpur, Malaysia

Mafas Raheem[2], V. Sivakumar[3]
Academic, School of Computing
Asia Pacific University of Technology & Innovation
Kuala Lumpur, Malaysia

***Abstract*—The revolution of big data has made resonance in the banking sector especially in dealing with the massive amount of data. The banks have the opportunity to know about the customer's opinions and satisfaction regarding their products by analyzing the data gathered every day. So, the banks can transform these data into high-quality information that allow banks to improve their business especially in credit cards which is becoming a short-term business for the banks nowadays. Further, the sentiment analysis has become immense in the field of data analytics especially the customers' opinion makes a huge impact in making profitable business decisions. The outcome of the sentiment analysis does assist the banks to know the deficiencies of their product and allow them to improve their products to satisfy the customers. From the sentiment analysis, 45% of the customers were negative, 30% were positive and 25% were neutral towards the credit card facility offered by the commercial banks. Also, the prediction of credit card customer satisfaction will contribute in a significant way to create new opportunities for the banks to enhance their promotion aspects as well as the credit card business in future. Random Forest algorithm was applied with three various experiments utilizing the normal data, balanced data and the optimized model with the normal data. The optimized model with the normal data obtained the highest accuracy of 87.38% followed by the normal dataset by 85.82% and the least accuracy was for the balanced dataset by 82.83%.**

Keywords—Credit card; predictive analytics; random forest; sentiment analysis; banking

I. Introduction

The credit card or "plastic money" is a digital money lending service offered by various financial firms which include banks and any other financial institutions. The main aim of this card is to allow its users to borrow money at the sales point to allow them to complete their purchases easily and conveniently. However, this convenience may result in problems for the users as the users should repay the banks the amount they borrowed along with the specified interest rates at the end [1].

Based on the study conducted on marketing tools, the banks try to acquire more customers to purchase credit card services [2]. The findings showed that credit card facility will attract more customers when the banks offer more incentives. This states that the incentives are an effective marketing supportive tool in convincing the customers to use a particular card in their specific purchasing aspects. Incentives such as cash rebates, lower interest, and airline miles may drive customers to spend more without thinking about the future [3]. Besides, another study stated that if the credit cards users are not fully aware of the incentives and the impact of their usage pattern of the credit cards might increase their future debt.

In the past few years, credit cards usage has been increased due to the incentives offered by financial institutions. At present, many credit cards users don't fully understand how those cards are operating. Yet, they are still using it just because of the offers they get from the cards. Nowadays people keep buying and exceeding their limits without recognizing that they are putting themselves in serious debt. However, if the credit cards are managed properly and users are fully aware of the pros and cons, it will be very beneficial for them in terms of its ease of use and convenience as it is the best alternative of cash [4].

In Malaysia, there are 3.6 million credit card users as of June 2017 and it was recorded that 18% of Malaysians between the age of 20 and 74 are using the credit cards based on the statistics provided by the Malaysian Department of statistics in 2016 [5]. In the meantime, the credit card outstanding balance was 36.9 billion Malaysian ringgits, where the overdue balance was 2.3 billion Malaysian ringgits which represent 7.3% [5]. It was also noted from the same source, the residual balance which has not exceeded 3 months of the repayment date was 2.3 billion Malaysian ringgits representing 6.2%, whereas the balance which has already exceeded 3 months was 0.4 billion Malaysian ringgits representing 1.1%. Furthermore, the statistics proved that in general, 43.6% of credit card users have paid their debts as well as another group of credit card users representing 43.6% settled leastwise 5% of their debts, whereas 12.8% of credit card users did not settle their debts. Besides, the aspect of insolvent cases, nearly 845 credit card users whose age is under 30 has declared their bankruptcy by the first half of 2017 [5].

This article is based on the study of credit card business in Malaysia and presented thematically in the following sections. Section 2 discusses literature regarding the relevant techniques and Section 3 covers the methodology followed in this study. Further, Sections 4 and 5 elaborates the descriptive and predictive analytics respectively which were conducted to offer actionable insights to the practitioners of the concerned domain.

II. Literature Review

A. Credit Card Business in Malaysia

The increase of the credit card users caused indiscriminate spending by the customers, which triggered various side effects as well. Numerous Malaysians happen to show excessive purchasing behaviour which is also known as "compulsive purchasing" [6]. Furthermore, it is considered easier for Malaysians to get a credit card rather than obtaining a loan facility due to the various complicated procedures. As a result of that, the millennials have adapted to move on with the modern lifestyle with the support of the credit cards. On the other hand, most of the credit card users have debts to settle which was due to the extensive use of the credit card facilities.

B. Bank Customers and Credit Cards

In general, consumers are inclined to recklessly using credit cards sometimes which ends up by overspending and having large amounts of debts. A study was conducted by interviewing fifteen young customers to grasp their perceptions and considerations regarding the use of credit cards. The results showed that there is a switch in spending behaviour. It was shown that the mind-set of the older generation is to save and spend later, whereas the mind-set of the younger generation is to borrow and spend now and pay later. The awareness of the advantages and disadvantages of the credit card facilities was also investigated in the same study. The results showed that young consumers are aware of the disadvantages which can cause a problem for them in the future, but they still use it and sometimes overspend due to the bonuses, points, discounts, gifts, and so on. Also, the world is moving toward a cashless society and the young customers are feeling satisfied with these cards as it offers them safety instead of carrying cash as well it offers them ease of use and convenience [7].

C. Factors Affecting Credit Card Usage

In Malaysia, numerous researchers have been examining the factors affecting the usage of credit cards which are influencing the customer’s spending patterns. For instance, it was stated from the study conducted in Malaysia, that the age has a considerable influence on credit card usage and their behaviour. Moreover, the level of income and material status has a significant effect on credit card usage where customers who have high income and married are likely to spend more using their credit cards compared to those who have low income and single. Also, this study considered the bank policies in which the findings showed that this factor has a significant influence where customers are likely to pay more with credit cards when they are offered with some benefits by the banks such as longer period and low annual interest rates [8].

Numerous studies have been done in many countries examining the usage of credit card and the customer’s behaviour. For instance, India as one of the biggest markets in the world, researchers have conducted studies on several factors such as gender, age, convenience, and sense of fulfilment. Results showed that convenience was the most influential where the ease of the use helps in increasing credit card usage. Also, based on the age factor, the young customers spend more with their credit cards compared to the aged customers who still prefer to use cash, and this had come into agreement with the studies conducted in Malaysia as well. Furthermore, the gender factor in India is influencing differently compared to the Malaysian context where in India the males are likely to spend more through their credit cards than females. The sense of fulfilment factor also has influenced the usage of credit card among Indians where individuals who have credit cards feel that it is an achievement in their lifestyle [9].

Similarly, a survey in Klang Valley area in Malaysia was conducted to examine the socio-economic and demographic factors including level of income, age, gender, and occupation towards the usage of the credit cards [10]. The results stated that gender does not influence the credit card usage, where the personal background of the individual has a considerable influence on usage. Also, this study mentioned that the credit card will be used for spending in various situations where some use it on basic things and others use it for retail purchases and proved that the customer profile has a significant relationship with the credit card usage [10] [11].

D. Sentiment Analysis in the Credit Card Business

Nowadays, managing the customers’ feedback has turned into a complicated activity due to the incorporation of various sources such as surveys on the customer's satisfaction, customers review and social media feedbacks. However, the capability to quickly grasp the feeling of the customers about the products is considered valuable as well as critical for businesses. Banks have globally started to utilize the data to obtain useful actionable knowledge via various ways such as sentiment analysis, reputational risk management, product cross-selling, financial crime management, and so on. At present, the banks can easily get benefited from big data where they can quickly extract high-quality information and then convert them into actionable knowledge which can improve the bank's performance [12].

Now-a-days, due to the advancements in technology the credit card is not used as in the past where people now are using e-wallets instead of the physical cards [13]. It was also mentioned that thousands of customers declared bankrupt due to the less controlled purchasing behaviour of the customers. However, some of the credit card facilities are considered as non-secured comparing to other types of debts so it is would be difficult for the banks to collect the debts if customers failed to pay the amount of debt and declared bankrupt [14].

In recent years, the amount of information is increasing due to the growing number of social media users at a rapid speed. Many businesses nowadays have no option other than analyzing the customer’s opinions and expressions regarding the products that they experienced. Furthermore, banks must know their customer's opinions on their products to formulate better marketing strategies to enjoy the competitive advantages among the imperfect competition prevailing in the market. Computerized sentiment analysis procedures provide more useful insights on the customer thoughts as well as a satisfaction to make effective actionable decisions. The sentiment analysis will help the banks to take immediate action to improve its services to the customers especially in the field of debit cards, credit cards and other online services [15].

E. Sentiment Analysis - Techniques and Algorithms

Sentiment analysis is performed in various domains to find out the opinions of the customer for better decision making. Supervised and Unsupervised Machine learning techniques are deployed in classifying and/or predicting the sentiments as positive or negative for the given opinion/reviews [16]. Many researchers have performed sentiment analysis to analyze the movie reviews to classify the sentiments into positive or negative using three different algorithms namely Maximum Entropy classifier, Support Vector Machine (SVM) and Naïve Bayes classifier [17]. Besides that, the three models were augmented using n-grams and the results have shown that the SVM model has outperformed the other two models.

Furthermore, a method based on Artificial Neural Network (ANN) was proposed to classify huge tweets dataset through Hadoop into positive or negative with the suggestion of fuzzy tone where the results were measured according to the accuracy and speed [18]. The results showed that the proposed model was very efficient in dealing with big sentiments datasets better than the small datasets. Also, the results showed that ANN has outperformed SVM and Hidden Markov Model (HMM).

Moreover, other researchers have suggested a new method for the classification of sentiments in the blogospheres and this proposed method was done through the combination of the advantages of Back-Propagation Neural Networks (BPN) and (SO) indexes. The results showed that the suggested method delivered more accurate outcomes and it was noticed that the classifying accuracy was improved with lesser training time [19].

Various libraries were utilized by researchers to perform sentiment analysis and Natural Language Toolkit (NLTK) and Textblob were the most famous. NLTK is considered as a platform that is utilized to build programs in Python and it works with the Natural Language data with statistical NLP application. Also, it includes text processing libraries for classification, tagging, parsing, and so on. Whereas Textblob is a library in Python that is structured on top of NLTK and utilized for NLP. Also, it is considered easier than NLTK as it possesses a simple API which is likely the easiest path to start with the sentiment analysis as well as various text analytics using Python [17].

F. Predictive Modelling in the Credit Card Business

So far, in the credit card market, all the researchers were focusing on credit risk as to the main problem as it is rising every day and has been a long-term issue for the banks. Therefore, various researchers used machine learning algorithms to build models with high prediction accuracy to predict the default rate of credit cardholders. Yet, there are not many pieces of research focusing mainly on the application of machine learning algorithms in predicting the credit card customer satisfaction as indeed a needed one for the current banking industry situation.

A study on credit card fraud detection using machine learning algorithms such as Logistic Regression, Random Forest, Naïve Bayes, and Multilayer Perceptron on the credit card fraud data with class balancing using SMOTE technique. The Random forest obtained the highest accuracy by 99.96% followed by MLP, NB, and LR by 99.93%, 99.23%, and 97.46%, respectively [20].

Another study was conducted on the credit card users default payment using six various machine learning models which are Regression Tree, Nearest Neighbors, SVM Regression, Random Forest Regression, Linear Regression and AdaBoost with the accuracies of 83%, 82%, 85%, 70%, 80%, and 88% respectively where the linear regression got the highest accuracy rate by 88% while the random forest got the lowest accuracy rate by 70% [21].

In the same domain of the credit default risk, six different machine learning models were built namely Naive Bayes, Bayesian networks, J48, random forest, multilayer perceptron, and logistic regression and evaluated using the accuracy, recall, and precision. The logistic regression obtained the highest accuracy (83.108%) whereas Naive Bayes, Bayesian networks, J48, Random forest, and Multilayer perceptron got 82.532%, 82.528%, 82.470%, 82.110%, 81.41%, respectively [22].

The prediction of the default of the credit card users was researched while having implemented balancing techniques such as SMOTE and ADASYN to balance the data [23]. It was noticed that the implemented models (SVM, KNN, Decision Tree and Random forest) achieved the highest accuracies with the normal dataset without any balancing techniques applied except the SVM model that showed the same performance in both normal and balanced data with ADASYN technique. SVM model accuracy showed a very small increase of 0.0025% after balancing the data with SMOTE technique which was considered not an improvement. The model's accuracies were 77.73%, 75.08%, 72.60%, and 80.88% respectively, and noted that the Random forest model achieved the highest accuracy rate of 80.88% with the normal dataset. However, the Recall and ROC values in both balancing techniques ADASYN and SMOTE of all the models were increased and showed values higher than the normal data.

III. Materials and Methods

A. Data Collection

The questionnaire is considered as a standard survey tool to collect data in quantitative research. A well-structured questionnaire was used to collect the required data for the research. Since the study population was already known (Young Professionals in Malaysia), the design of the questionnaire was straight-forward and easy with all related instructions. Besides, the questionnaire included both open- and closed-ended questions to minimize the bias and to increase the respondent’s rate.

B. Data Analysis

R studio and Tableau were chosen to perform the analytics to attain the specified objectives of this study. Feature selection is defined as a pre-processing strategy that can be effective in the preparation of data for the various problems in machine learning and data mining. Therefore, the regression and the Correlation-based Feature Selection (CFS) methods were used to choose the most influential attributes. This strategy found useful in acquiring clearer and simple models, with enhanced performance of data mining [24].

The data was pre-processed via various steps such as transformation, cleaning the missing values and detecting outliers if they exist. After cleaning the dataset, descriptive analytics was conducted in which suitable visualizations were used to understand the variables and their relations which revealed some useful insights regarding the scenario. Besides that, after understanding the dataset, predictive analytics particularly sentiment analysis was conducted to figure out the different patterns and opinions of customers regarding the credit card business as well as a predictive model was built to predict the credit card customer satisfaction.

C. Balancing Techniques

A predictive model building using machine learning algorithms is significantly affected by imbalance data. The target variable got two classes (YES and NO) where (NO) was the majority. Therefore, to obtain fair results for the minority class while building the predictive model, class balancing techniques were applied, and this led to balance the dataset distribution.

Three various resampling techniques are generally applied in the class balancing such as over-sampling, under-sampling and hybrid. However, the over-sampling technique was applied over the other techniques due to the less number of records exist in the data set. The variance among oversampling and under-sampling is shown in Fig. 1.

D. Classification Algorithm - Random Forest (RF)

The Random Forest algorithm is a tree-based model in which it is built through the combination of the predictions of different trees where each tree is trained individually [26]. RF model is considered as one of the important competitors to the latest algorithms such as SVM and Boosting. Besides, what differentiates the RF model from the other models is that it is simple as well as fast to apply. Also, it is known for its prominent performance where it gives high accurate predictions. The random forest algorithm can deal with a large number of variables. With each decision split in the RF model, the variables are selected randomly and as a result, this drives the correlation among the trees to decrease [27]. Thus, the power of prediction will be enhanced and as a result, it ends up with higher model efficiency.

Additionally, the RF can assist in extracting the most effective variables that have an impact on the target variables. As a consequence of that, applying this model for predicting the credit card customer satisfaction is substantial. However, the explanation of the RF model is difficult when contrasted to a single decision tree. Therefore, to make the RF model more understandable, there are some possible ways one of them is the estimation of the importance of the features.

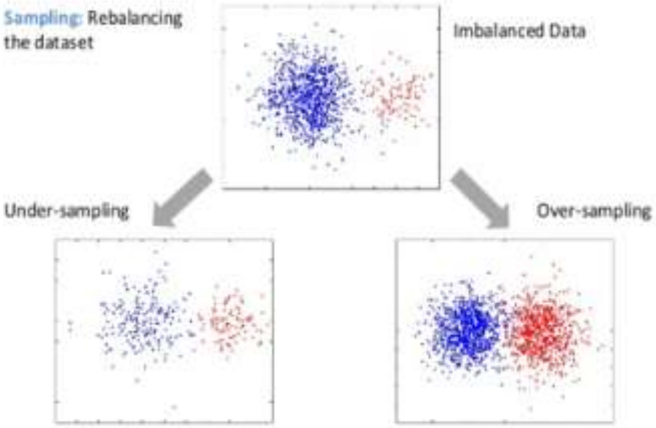


Fig. 1. Over-Sampling and Under-Sampling Methods [25].

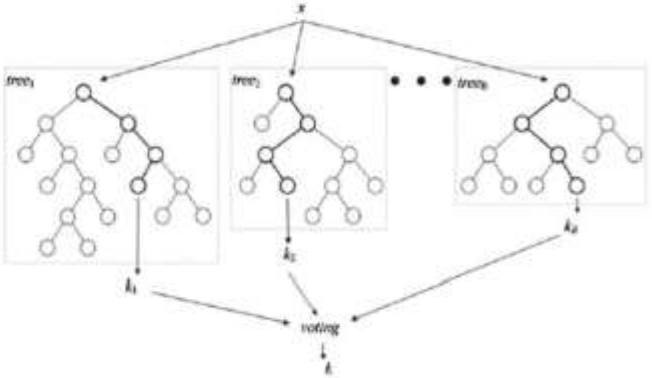


Fig. 2. Random Forest Trees [28].

A tuning parameter named "mtry" was utilized to optimize the RF model performance in which its task falls short in presenting the variables number sampled randomly as candidates within every division of the decision tree as depicted in Fig. 2 [28].

E. Conceptual Framework

As shown in Fig. 3, a conceptual framework is considered as a tool to determine the potential relationships among the independent and dependent variables. Also, it is an incorporated technique to consider the problem of the research with the most suitable approach to the study [29].

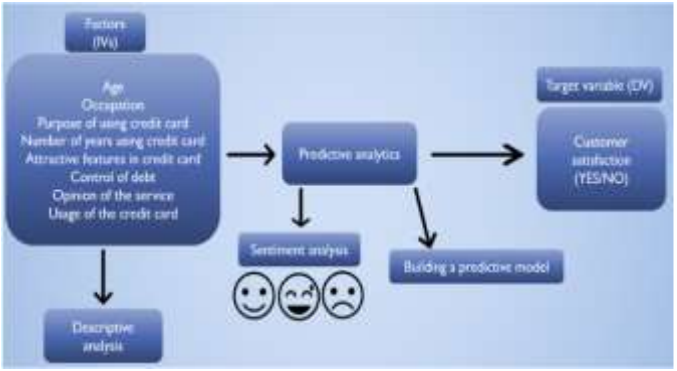


Fig. 3. Conceptual Framework.

IV. Descriptive Analytics

The main focus was on one of the most important variables which impact on the behaviour of the customers when using their credit cards. The usage of the credit card variables got five classes in which the promotion class has been selected by 283 individuals. As depicted in Fig. 4, 163 males and 120 females were found who spend using their credit cards under the promotions. Besides, 187 were employed where the remaining 96 were not employed from the 283 individuals. When it comes to the education of this specific group, it was seen that 32 diploma, 22 secondary, 152 were holding a bachelor's degree, 63 masters' degree, and 14 doctorates.

Also, it was clear that the age range 18-24 and 25-34 were representing the majority of this group while the minority was represented by the age range 35 and above. In terms of the satisfaction of the customer, it was seen that from the 283 individuals who were spending with their credit cards during promotions, a total of 222 were not satisfied with the service while only 61 of them were satisfied. Therefore, it could be concluded that the previously analyzed variables were connected and having an impact on the target variable.

So, it could be concluded that young people who were between the age of 18-24 and 25-34 got influenced more than elder people by promotion which drives them to spend more recklessly without thinking about debt problems. Also, the education level got an impact as most of them were bachelor's degree holders with less information about credit card facilities and led them to be dissatisfied.

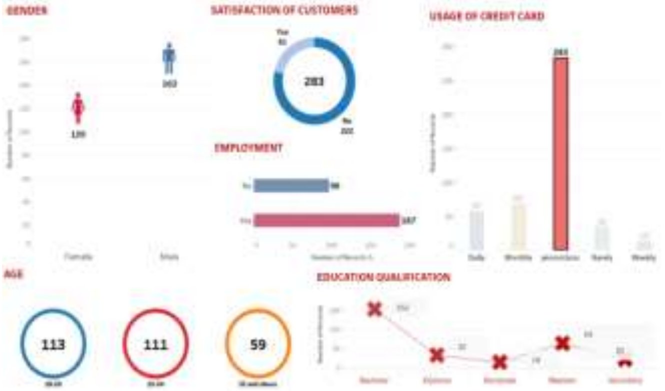


Fig. 4. Data Visualisation 1.

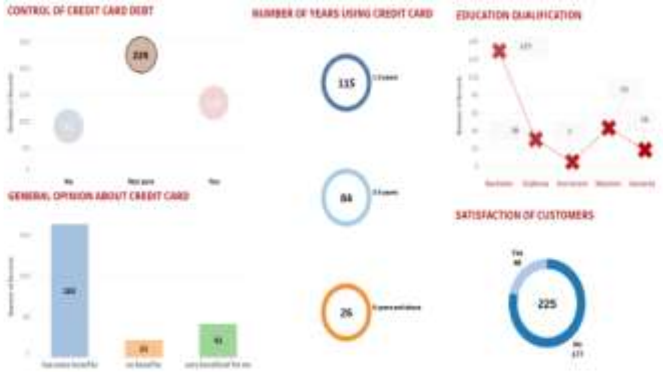


Fig. 5. Data Visualisation 2.

Fig. 5 mainly focuses on the control of the customers on their credit card debts. It was noticed that a large portion of the respondents was not sure of whether they can control their credit card debts or not. Also, it can be seen that the majority of those people who were not sure belong to the groups that have been using the credit card for 1-2 years (115 individuals) and the group of 3-5 years (84 individual) while the minority of 26 have been using it for 6 years and above. In terms of their education, the bachelors and masters got the majority as 129 and 43 individuals, respectively. Besides, their general opinion about the credit cards with benefits for them represented by 163 individuals while remaining selected no benefits and very beneficial for me by 21 and 41 respectively. When it comes to the target variable, from the 225 individuals 177 were not satisfied with the service compared to 48 who were satisfied. It can be concluded that customers who were using credit cards for short periods mostly holding bachelors' degree and were using it for getting some benefits by spending without thinking about future problems.

V. Predictive Analytics

A. Modelling: Random Forest Model (RF)

Random Forest is a machine learning algorithm which creates several classification trees through utilizing the bootstrap sampling technique while training the model and, the classification trees produce the final prediction in the test phase. This algorithm was applied to both the unbalanced and balanced dataset and it gave a better performance on the unbalanced dataset.

Fig. 6 displays the Out-of-bag (OOB) error plot which is utilized to measure the prediction error rate of the Random Forest model as well as it identifies the number of trees utilized during the model building process. Also, the error rate used to reduce with the increase in the number of trees. From the three lines, the black line shows the overall error rate, while the green and red lines are referring to one class error, respectively.

1) Imbalanced dataset: It can be seen that the increased rate of error decreased when it became stable in all three lines, which means that after 200-300 trees, the error rate did not decrease any more.

Fig. 7 represents the RF model performance with the accuracy of 85.82%.

2) Balanced dataset (oversampling technique): Fig. 8 shows the balanced data using the oversampling technique where both classes (Yes and No) got balanced with 50% each and the RF model obtained 82.83% accuracy.

3) Random forest model optimization: The RF model was tuned using the hyperparameters among which the mtry was very prominent. The task of this (mtry) is to represent the features number randomly within every node of the RF tree. Fig. 9 shows that when the mtry is 6 the model got the lowest error rate of 12.62% compared to the mtry 12 and mtry 3 with an error rate of 12.93% and 13.56% respectively. Thus, mtry 6 was selected as the lowest error rate which gave an accuracy of 87.38%.

Random Forest algorithm list out the most important variables while building the classification model and this is considered as one of the most significant outcomes of this algorithm as shown in Fig. 10.

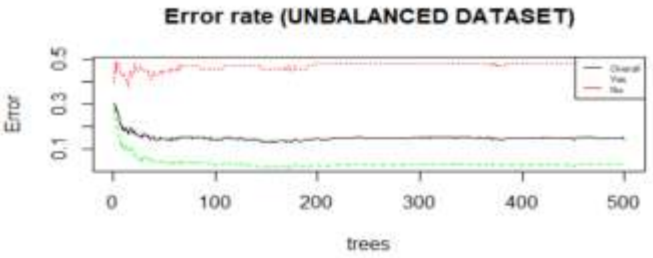
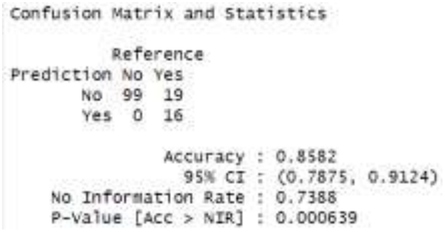


Fig. 6. Unbalanced Dataset Error Rate.



```
Confusion Matrix and Statistics

          Reference
Prediction No Yes
       No  99  19
       Yes  0  16

               Accuracy : 0.8582
                 95% CI : (0.7875, 0.9124)
    No Information Rate : 0.7388
    P-Value [Acc > NIR] : 0.000639
```

Fig. 7. RF Confusion Matrix for Imbalanced Dataset.

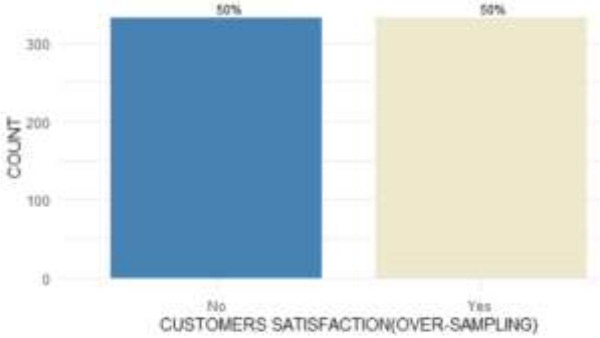


Fig. 8. Target Variable after Balancing.

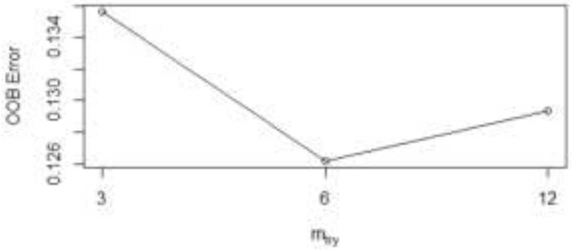


Fig. 9. Error Rate Plot.

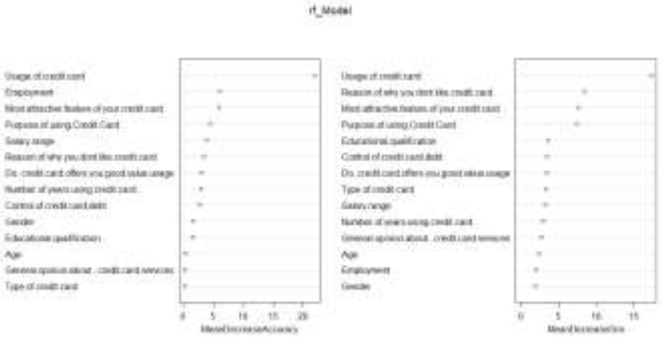


Fig. 10. Factors affecting the Credit Card User's Satisfaction.

The significance of the variables is arranged from top to bottom in which the top variables are with the most significant. Also, there are two kinds of the mean within the figure above which are mean decrease accuracy and mean decrease Gini where the first type represents the most important variable to the least significant by sorting them from top to bottom and the second type represents the measurements of every variable contribution to the similarity of leaves and nodes in the RF model. Besides, it was noted that the usage of the credit card variable got the most effect on the credit card customer satisfaction, while Types of the credit card variable got the least effect.

4) RF Model experiments: Table I and Fig. 11 shows all the information's regarding the RF model experiments which state the model accuracy, error rate, recall, and precision percentages.

It could be seen that the optimized model through mtry hyper-parameter achieved the highest accuracy rate of 87.38% followed by the normal dataset with 85.82% and the lowest accuracy rate obtained for the balanced dataset of 82.83%. It is known that the higher recall and precision rates indicate better performance of the model. The precision value for the normal dataset was 100%, where the balanced dataset and the optimized model got 82.75% each. Furthermore, the recall rate for the normal data was 83.89% while it was 82.75% each for the balanced dataset and optimized model. Lastly, the optimized model achieved the lowest error rate of 12.62% followed by the normal data with 14.18% and the highest error rate of 17.17% was achieved for the balanced dataset.

TABLE I. RANDOM FOREST MODEL COMPARISON

Data	**Fine-tuning**	**Accuracy (%)**	**Recall (%)**	**Precision (%)**
Normal Data(Imbalanced)	-	85.82	83.89	100
Balanced Data(Over-sampling)	-	82.83	82.75	96.96
Optimized Model with Normal Data	Mtry	87.38	82.75	96.96

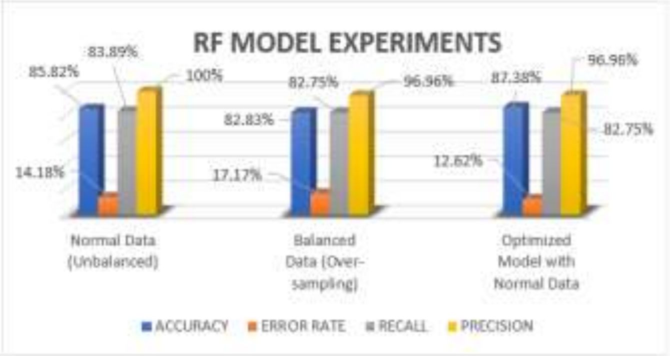


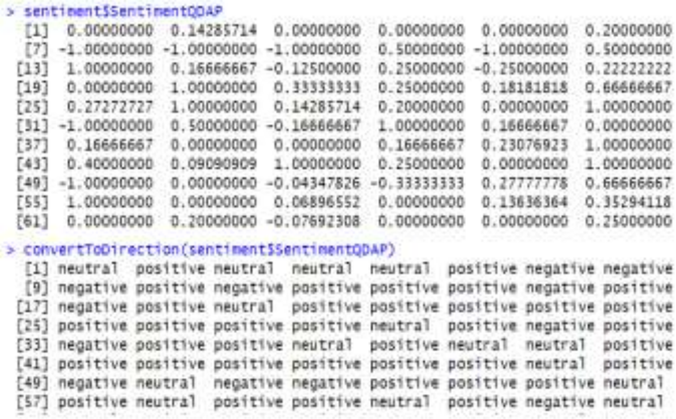
Fig. 11. RF Model Performance.

B. Sentiment Analysis

The Specific opinion of the customer about the credit card variable was used to perform the sentiment analysis using various opinions of the customers. The sentiments of the customers were predicted as positive, neutral, and negative which were represented by 1, 0, and -1, respectively. Fig. 12 shows the predicted sentiments according to the QDAP dictionary and it was presented by 1, 0, and -1 with decimals according to the unsupervised rule-based prediction method.

Fig. 13 displays the sentiment plot where it was noticed that negative sentiments were dominant over positive and neutral sentiments. Also, it was noticed from the opinions of the customers that even if the sentence is positive there was negative comment within the sentence itself, for example, “I like the credit card service and it helps me in some situations, however, I will always be worried from fraud and overspending”. Almost more than half of the data was in this form where it starts with positive words then stating the negative opinion which caused dissatisfaction with the credit card service. This was the main reason to get more negative compared to positive and neutral sentiments.

The overall sentiments were represented using a pie chart as shown in Fig. 14. The negative sentiments got the highest percentage of 45% followed by positive sentiment (30%), and the least 25% for neutral sentiments.



```
> sentiment$SentimentQDAP
 [1]  0.00000000  0.14285714  0.00000000  0.00000000  0.00000000  0.20000000
 [7] -1.00000000 -1.00000000 -1.00000000  0.50000000 -1.00000000  0.50000000
[13]  1.00000000  0.16666667 -0.12500000  0.25000000 -0.25000000  0.22222222
[19]  0.00000000  1.00000000  0.33333333  0.25000000  0.18181818  0.66666667
[25]  0.27272727  1.00000000  0.14285714  0.20000000  0.00000000  1.00000000
[31] -1.00000000  0.50000000 -0.16666667  1.00000000  0.16666667  0.00000000
[37]  0.16666667  0.00000000  0.00000000  0.16666667  0.23076923  1.00000000
[43]  0.40000000  0.09090909  1.00000000  0.25000000  0.00000000  1.00000000
[49] -1.00000000  0.00000000 -0.04347826 -0.33333333  0.27777778  0.66666667
[55]  1.00000000  0.00000000  0.06896552  0.00000000  0.13636364  0.35294118
[61]  0.00000000  0.20000000 -0.07692308  0.00000000  0.00000000  0.25000000
> convertToDirection(sentiment$SentimentQDAP)
 [1] neutral  positive neutral  neutral  neutral  positive negative negative
 [9] negative positive negative positive positive positive negative positive
[17] negative positive neutral  positive positive positive positive positive
[25] positive positive positive positive neutral  positive negative positive
[33] negative positive positive neutral  positive neutral  neutral  positive
[41] positive positive positive positive positive positive neutral  positive
[49] negative neutral  negative negative positive positive positive neutral
[57] positive neutral  positive positive neutral  positive negative neutral
```

Fig. 12. Sentiment Classification.

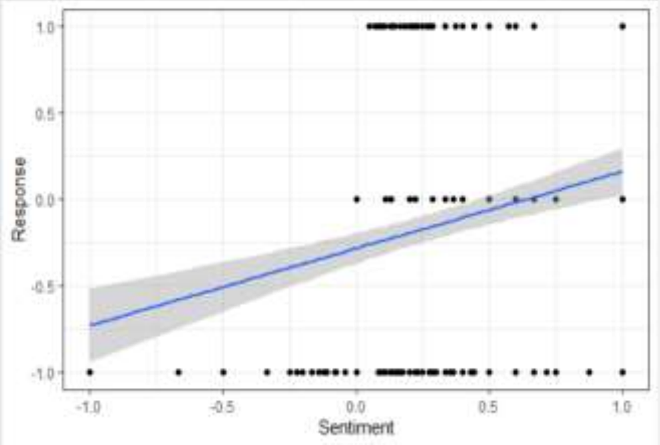


Fig. 13. Sentiment Analysis Plot.

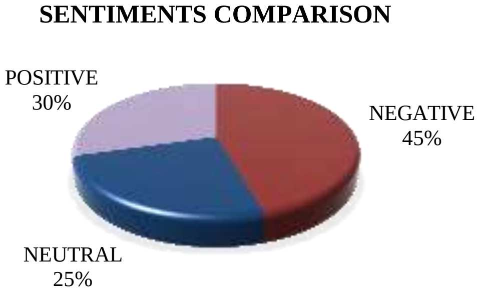


Fig. 14. Sentiments Comparison.

According to the results obtained from the sentiment analysis, it can be seen that most of the users were dissatisfied and had negative opinions about the credit card services which may indicate the existence of other problems such as customers who are not able to carry out their responsibilities with their credit card spending habits and it ends up with debts and financial problems. Also, it may indicate that some customers are not financially literate, and they keep spending more money without realizing that they put themselves in difficult situations that drive them to damage their financial future while the banks are creating more profits from the interests charged to the credit cards.

VI. Conclusion

The main aim of this study was to perform predictive analytics particularly sentiment analysis as well as building a predictive model to predict customers’ satisfaction on the credit card services. The data mining and data visualization were utilized to perform all the tasks as well as to explore all the variables and show the effect on the credit card customer satisfaction. The oversampling technique was applied to perform class balancing and a predictive model was built using the Random Forest algorithm. Also, three different experiments were conducted using normal data, balanced data, and optimized model with normal data to achieve the objectives of this study.

As stated earlier, the best accuracy of 87.38% was obtained for the optimized model through the mtry hyper-parameter. On the other hand, sentiment analysis on the specific opinion of the customers on credit card services was performed and found that 45% of the responses reflected negative sentiments. It was concluded that the people in the age range of 18 to 34 were explicitly utilizing the credit card and generally fail to pay back their debts or always worried of fraud and risk of spending more carelessly. The outcome obtained from this study would be more useful to the decision-makers in the banking sector concerning the development of the credit card business in future.

There can still be improvements in applying robust machine learning algorithms and/or deep learning architecture with proper optimization techniques to build a more effective classification model. Also, socio-economic variables can be included in the analytics process to get a high-quality insight which would be more useful to the decision-makers. However, limited time and restricted access to data due to the current pandemic situation were the major limitations in this study. So, increasing the sample size of the data as well as adding new variables is recommended to obtain better results in all means in future.

REFERENCES

[1] C. Bertaut, C. and Haliassos, M. “Credit Cards: Facts and Theories,” SSRN Electronic Journal, 2005, doi:10.2139/ssrn.931179.

[2] Agarwal, S. “Why Do Banks Reward Their Customers to Use Their Credit Cards?,” 2011, FRB of Chicago Working Paper.

[3] Chatterjee, P. & Rose, R. “Do Payment Mechanisms Change the Way Consumers Perceive Products,” Journal of Consumer Research, 38(6), 2012, pp.1129-1139, doi: 10.1086/661730.

[4] E. Halim, R., Adiwijaya, K., Haryanto, J. and Firmanzah, F. “The Propensity of Young Consumers to Overspend on Credit Cards: Decomposition Effect in the Theory of Planned Behavior,” Journal of Economics, Business and Management, 4(10), 2016, doi: 10.18178/joebm.2016.4.10.459.

[5] The sun daily. “3.6m credit card holders in Malaysia as at June,” [online] p.A single page. Available at: https://www.thesundaily.my/archive/36m-credit-card-holders-malaysia-june-ITARCH473259, 2017, [Accessed 4 Jul. 2020].

[6] UKEssays. “Overview of Credit Cards In Malaysia,” Marketing Essay. [online]. Available from: https://www.ukessays.com/essays/marketing/overview-of-credit-cards-in-malaysia-marketing-essay.php?vref=1, November 2018, [Accessed 14 December 2020].

[7] Lim, W. M., Ng, W. K., Chin, J. H. and Boo, A. W. X. “Understanding Young Consumer Perceptions on Credit Card Usage: Implications for Responsible Consumption,” Contemporary Management Research, 10(4), 2014, doi:10.7903/cmr.11657.

[8] Teoh, W. M. Y., Chong, S. C. and Yong, S. M. “Exploring the factors influencing credit card spending behavior among Malaysians,” International Journal of Bank Marketing, 31(6), 2013, DOI: 10.1108/IJBM-04-2013-0037.

[9] Khare, A., Khare, A. and Singh, S. “Factors affecting credit card use in India,” Asia Pacific Journal of Marketing and Logistics, 24(2), 2012, pp. 236-256, doi: 10.1108/13555851211218048.

[10] Hussin, S. R., Kassim, S., and Jamal, N. “Credit Card Holders in Malaysia: Customer Characteristics and Credit Card Usage,” Int. Journal of Economics and Management, 7(1), 2013, pp.108 – 122.

[11] Ahmed, Z. U., Ismail, I., Sadiq Sohail, M., Tabsh, I., and Alias, H. “Malaysian consumers’ credit card usage behaviour,” Asia Pacific Journal of Marketing and Logistics, 22(4), 2010, pp.528–544. doi:10.1108/13555851011090547.

[12] Srivastava, U. and Gopalkrishnan, S. “Impact of Big Data Analytics on Banking Sector: Learning for Indian Banks,” 2nd International Symposium on Big Data and Cloud Computing (ISBCC’15), 2015, pp.643–652. doi: 10.1016/j.procs.2015.04.098.

[13] New Straits Times. “Number of bankrupts due to credit card usage relatively low,” [online] Available at: https://www.nst.com.my/business/2018/03/345048/number-bankrupts-due-credit-card-usage-relatively-low, 2019, [Accessed 12 Jul. 2020].

[14] Ismail, S. “Watch out for credit card debt. New Straits Times,” [online]. Available at: https://www.nst.com.my/opinion/letters/2019/02/462125/watch-out-credit-card-debt, 2019, [Accessed 1 Jul. 2020].

[15] N.Sumathi and T.Sheela. "Opinion Mining Analysis in Banking System using Rough Feature Selection Technique from Social Media Text," International Journal of Mechanical Engineering and Technology (IJMET), 8(12), 2017, pp.274–289.

[16] Borele, P. and Borikar, D. A. “An Approach to Sentiment Analysis using Artificial Neural Network with Comparative Analysis of Different Techniques,” Journal of Computer Engineering (IOSR-JCE), 18(2), 2016, pp.64-69.

[17] Pang, B., Lee, L. and Vaithyanathan, S. “Thumbs up? Sentiment Classification using Machine Learning Techniques,” 2002, Philadelphia, s.n.

[18] Jian, Z., Chen, X. and Han-shi, W. “Sentiment classification using the theory of ANNs,” The Journal of China Universities of Posts and Telecommunications, Volume 17, 2010, pp.58-62.

[19] Chena, L. S., Liub, C. H. and Chiu, H.-J. “A neural network based approach for sentiment classification in the blogosphere,” Journal of Informetrics, 5(2), 2011, pp.313-322.

[20] Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., and Anderla, A. “Credit Card Fraud Detection - Machine Learning methods,” INFOTEH-JAHORINA, 2019, 18th International Symposium.

[21] Ullah, M. A., Alam, M. M., Sultana, S. and Toma, R. S. “Predicting Default Payment of Credit Card Users: Applying Data Mining Techniques 2018,” 2nd Int. Conf. on Innovations in Science, Engineering and Technology (ICISET), 2018, Chittagong, Bangladesh, s.n.

[22] Çıgsar, B. and Unal, D. “Comparison of Data Mining Classification Algorithms Determining the Default Risk,” Scientific Programming, 2019, pp.1-8.

[23] Shetty, A. S. and R, M. “Prediction Of Default Credit Card Users Using Data Mining Techniques,” International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8(7), 2019, pp.2278-3075.

[24] Jundong, L., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J. and Liu, H. “Feature Selection: A Data Perspective,” ACM Computing Surveys (CSUR), 50(6), 2016, pp.1–44. doi: 10.1145/3136625.

[25] Walimbe, R. “Handling Imbalanced Dataset in Supervised Learning Using Family of SMOTE Algorithm,” Data Science Central. Available at: https://www.datasciencecentral.com/profiles/blogs/handling-imbalanced-data-sets-in-supervised-learning-using-family, 2017, (Accessed: 31 August 2019).

[26] Denil, M., Matheson, D. and Freitas, N. De. 2014. Narrowing the Gap : Random Forests In Theory and In Practice, in Proceedings of the 31st International Conference on Machine Learning. Beijing, China: JMLR, pp. 1–9. Available at: http://proceedings.mlr.press/v32/denil14.pdf.

[27] Biau, G. “Analysis of a Random Forests Model,” Journal of Machine Learning Research, 13(1), 2012, pp.1063–1095. Available at: http://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf.

[28] Esteves, G. and Mendes-Moreira, J. “Churn Prediction in the Telecom Business,” Eleventh International Conference on Digital Information Management (ICDIM). Porto, Portugal: IEEE, 2016, pp.254–259. doi: 10.1109/ICDIM.2016.7829775.

[29] Liehr, P., Smith M. J. “Middle range theory: Spinning research and practice to create knowledge for the new millennium,” Advances in Nursing Science, 21(4), 199, pp.81-91.

Automatic Detection of Elbow Abnormalities in X-ray Imagery

Mashal Afzal[1], M. Moazzam Jawaid[2], Rizwan Badar Baloch[3], Sanam Narejo[4]

Department of Computer System Engineering
Mehran University of Engineering and Technology
Jamshoro, Pakistan

Abstract—**Abnormality or deformity in any of the bone disrupts overall function of human skeleton. Hence, orthopedic abnormalities are common reasons for emergency department visits and elbow deformation is one of the common issue seen among emergency patients. Despite high frequency of elbow-related casualties, there is no standardized method for interpretation of digital X-rays. Accordingly, we have proposed a model for automatic deformation detection in elbow and connected forearm bones using Image Processing techniques. The X-ray images were preprocessed and the region of interest is segmented using Multi Class Probabilistic Segmentation in first step. Subsequently, multi-phase canny edge detector was used to highlight the edges and descriptive features were extracted to differentiate among normal and abnormal X-rays. On the basis of those features, three tests were performed to automatically trace deformities in different bones associated with elbow. The publically available Musculoskeletal Radiographs (MURA) dataset has been used in this research. Hence, 250 elbow X-rays from the dataset were investigated for geometrical shape distortions, crack, damage and extra-ordinary distance between the bones. Accordingly, the proposed method shows promising results in terms of 86.20% accuracy.**

***Keywords**—Deformation detection; multi-class probabilistic segmentation; edge detection and geometrical shape detection*

I. Introduction

In human skeleton which comprises 207 bones, every bone attains a very stimulating role and has same importance in entirety of skeleton. A minute abnormality or deformation in any of the bone can influence general structure of human skeleton.

The abnormality can occur in any bone such as leg, finger, femur, hand, rib cage, shoulder, chest, humerus and knee, etc. From the recent study, it is apparent that orthopedic abnormalities are one of the prevalent reason for emergency department visits [1]. Among these visits, elbow associated causalities are one of the common issues for the emergency visits. Some of the foremost reasons of elbow abnormalities are weight lifting, road accident, job involving repetitive hand moment, sports, trapped nerves, physical trainings, stiffness and arthritis.

In adults, the ration of elbow and forearm fractures is around eight to ten percent which are increasing steadily [2]. Over and above 1.7 billion people worldwide are effected with orthopedic abnormalities, which results in long-lasting, unbearable pain and disorders in human body [3].

For this growing rate of abnormalities, Medical image diagnosing tools are indispensable in this age [4]. One of the common pattern for abnormality diagnosing is X-ray which gives the shadow-like image as a result. Comparing with other image diagnosing methods, X-ray images are effective to use due to easy availability, inexpensiveness, harmlessness and high speed [5].

The digitization of medical imagery is an imperative trend now days. Despite high regularity of elbow-associated abnormalities, there isn’t any standardized method available for interpretation of digital X-ray images [6]. This research, emphasis on digitizing deformation detection for elbow X-ray images Fig. 1.

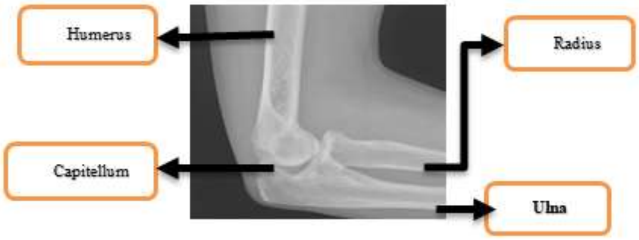


Fig. 1. Lateral View of Elbow X-ray Referring to Different Bones Present in Human Elbow [12].

Deformation or abnormality in elbow bone occurs in many ways depending on the number of joints and nature of the bone [7]. In this study, we will focus on categorizing elbow X-ray image as normal or abnormal by following the methodology discussed below in Section 3.

The work flow for this study comprises four more sections. The related work to diagnose abnormality in the bones of elbow and to wisely interpret X-ray images has been concisely presented in Section 2. Later on, in next two sections (3 and 4) we have presented the methodology followed to achieve our goal and the results based on numerous experiments. Furthermore, the conclusion of our study and the future work is presented in next section.

II. Literature Review

Elbow abnormalities can be detected through many ways using medical imaging but due to easy availability of X-ray images, this technique is widely used and suggested by the orthopedics [5]. Earlier, the deformation detection in elbow X-ray images has been performed manually but with the rise in technology some methods have been proposed to digitally interpret the X-ray images which are discussed below.

The deformation in elbow X-ray images depends on the nature of bone as there are four bones involved in human elbow. The two bones radius and ulna which runs from forearm to elbow joint through capitellum which further joints these two bones with humerus bone that runs from elbow to shoulder. So, the deformation in any bone involved in elbow can effect multiple other bones too as they all are interconnected to form the skeleton.

In [8], the author presents the idea to detect abnormalities in capitellum, it is the major joint in between the other three bones involved in elbow. To resist the abnormality in other bones the deformation in capitellum needs to be diagnosed at initial stage. This study states that the capitellum fractures are minor and mostly ignored at emergency visits but these minor fractures cause major abnormality, if not treated properly. To detect the capitellum abnormality, the author suggests the idea to analyze the arc sign around circular shaped capitellum. The arc sign is the key point to differentiate between normal and deformed elbow X-ray. All the deformed X-ray images which include arc sign were further treated. After the treatment, the arc sign was not seen in the usual X-ray images. A very basic concept was suggested by the author however; this analysis was performed manually which didn’t resolve the issue of digitally interpreting elbow X-ray images.

In [9], the author presents the idea to diagnose abnormality in humerus bone of elbow. A long bone running from shoulder to elbow with one end joined with capitellum. The ends of long bones are covered through a rubber type padding in human body called as cartilage. The discontinuity in between two cartilages at the ends of humerus bones direct towards geometrical shape deformation in elbow. To cover this problem, the author used canny edge detection method on humerus X-ray images to find sharp boundaries of humerus bone. On the resulting image, the parametric equation of straight line is used to detect straight line at the boundaries of humerus bone. The discontinuity or an angular curve between the lines was analyzed as geometrical shape deformation in humerus bone of elbow. This method was implemented on 48 X-ray images with 79.3% accuracy, however it only works for separate humerus X-rays and do not work if the other bones are involved in elbow X-ray image.

Another study has been proposed for the detection of fracture in radius and ulna bones using image segmentation techniques [10]. This study works for finding fracture in several long bones including femur, leg, radius and ulna. This study suggests to apply wiener filter to remove noise from images and then smoothing the images by using cumulative and probability distribution. Subsequently, image segmentation has been applied to get the region of interest which is followed by Edge detection method. Moreover, this study uses Hough transform to detect the straight lines to detect the fracture in similar manner as the author analyzed the fracture in previous study [9]. This algorithm detects fracture with 89.6% accuracy but the drawback for this study is similar as the previous that it only detects fracture in horizontal poses of X-ray images for different bones. Whenever the radius or ulna bone is linked with capitellum, the algorithm fails to detect any deformation.

Another interesting study has been presented by the author which detects radial head fracture at the top of radius bone at the main joint capitellum [11]. Normally, there’s a small rise at the joint when the X-ray is normal but in case of abnormality the rise disappears or appears geometrically disturbed. The author has presented very meaningful idea but this was also practiced manually on elbow X-ray images.

The main problem of the manual analysis of X-ray images could be solved by automating the fracture detection methods. Besides these studies, there are many other studies to interpret the elbow abnormalities by using different image modalities. Our study focuses on automating the interpretation of elbow X-ray images using certain techniques discussed in next section.

III. Methodology

In this portion, the proposed methodology is discussed to automatically detect the abnormality or any deformation in any bone of elbow. Three tests are proposed to check the elbow abnormality. If the X-ray image clears all the tests, it will be considered as normal X-ray image otherwise it will be classified as abnormal X-ray. The flowchart for the proposed approach is given below in Fig. 2.

A. Data Acquirement and Pre-Processing

The dataset used for the study is assembled from the website of Stanford ML group. Musculoskeletal Radiographs (MURA) is released by Stanford ML group [12]. This dataset contains X-ray images of different bones including elbow X-rays, hand X-rays, shoulder X-rays and finger X-rays, etc. 250 elbow X-rays are used for this study from the dataset including 190 normal and 60 abnormal X-rays. The X-ray images are preprocessed using image enhancement techniques and then resized and converted into gray scale images and finally processed for the next step.

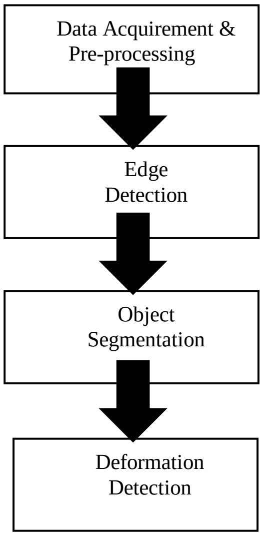


Fig. 2. Flowchart of Proposed Methodology.

B. Object Segmentation

The X-ray images contain background, flash and bone. To get the region of interest i.e. elbow bone, multi class probabilistic segmentation is applied on the preprocessed X-ray images [13]. Multi class probabilistic segmentation works by computing prior probabilities for each pixel of image and discovers the evidence of the similarity of the pixel with the next one. On the basis of the evidence, the posterior probabilities are calculated and the pixels are divided into multiple classes. The result of three class probabilistic segmentation as there are three regions in the images is shown in Fig. 3.

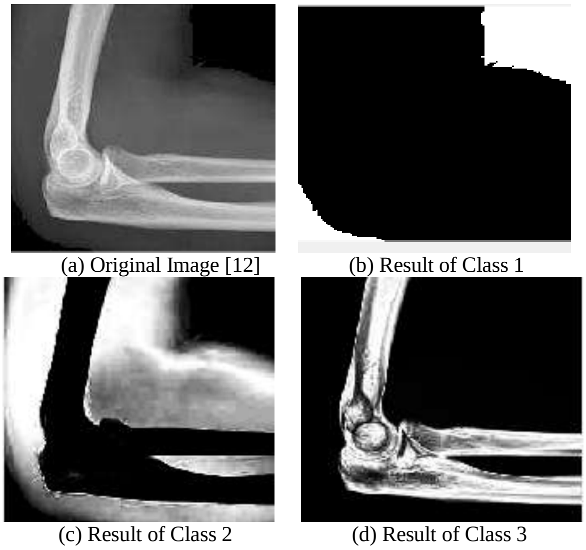
For class 1, all the pixels for background consuming same values are enhanced as shown in Fig. 3(b). For the remaining area, the posterior probabilities are recomputed and the result of class 2 is shown below in Fig. 3(c).

As a result of class 3, the pixels having same values for bone are highlighted as shown in Fig. 3(d). With the help of three class probablistic segmentation, the bone is extracted from the X-ray image and further processed in next step.

C. Edge Detection

On the segmented image, canny edge detection method is applied. The multi-phase edge detector is applied to highlight wide-ranging edges and boundaries on the resulting X-ray [14]. With the help of canny method, the boundaries of the elbow bone are highlighted as shown below in Fig. 4(b).

The edge highlighted image is further processed in next step where several features are extracted to differentiate between normal and abnormal elbow X-ray image.


(a) Original Image [12] (b) Result of Class 1

(c) Result of Class 2 (d) Result of Class 3

Fig. 3. Probablistis Segmentation for Three Regions. (a) Indicates the Original Elbow X-ray Image. (b) Indicates the Result for Segmentation of First Class which Highlights the Pixels of Background. (c) Shows the Result of Second Class Segmentation in which the Pixels Involved in Flash are highlighted. (d) shows the region of interest as a result of third class Segmentation in which the Bone in the X-ray Image is Highlighted.

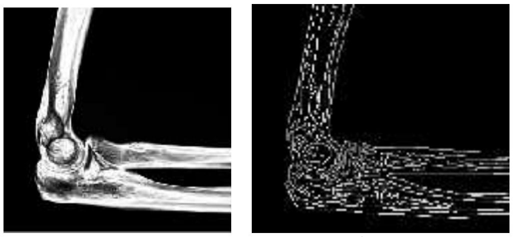

(a) Segmented ROI. (b) Mask of Canny Method.

Fig. 4. Edge Perceived Bone. (a) Indicates the Segmented Bone from Actual X-ray Image. (b) Shows the Edge Detected Image through Canny Method.

D. Deformation Detection

The deformation in any bone of elbow is highlighted in this step. So, this step is very imperative among all steps of methodology. On the basis of possible reasons of deformation in elbow X-ray that has been learnt from literature, three tests are suggested to classify the images to normal and abnormal classes. The three suggested tests to differentiate in between normal and abnormal X-rays are enlisted below.

- Test: 01 for Capitellum Deformation Detection.
- Test: 02 for Radial or Ulnar Fracture Recognition.
- Test: 03 for Geometrical Shape Abnormality Identification.

1) Capitellum deformation detection: The very first test is related with capitellum that is the smooth, rounded eminence and major joint which holds overall structure of elbow, the minor deformation at this joint will effect overall erection of elbow [15]. To detect deformation in capitellum, Circular Hough transform is applied to identify the circular shaped capitellum [16]. CHT detects the capitellum using the range of pixels for radius as 16 ± 2 pixels for this test. This range is acknowledged by practicing on various normal X-rays in which capitellum is not dislocated or distorted. The result of this process is shown below in Fig. 5.

The capitellum is located back on the original X-ray image as illustrated in 5(b). All the elbow X-rays which intent capitellum will be categorized as normal at this stage and further processed through other tests. All those X-rays which didn't reflect capitellum will be categorized as abnormal as the deformation or dislocation of capitellum is a major sign of abnormality.

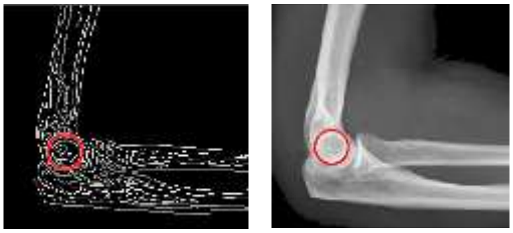

(a) Mask of Canny Method (b) Capitellum located Image

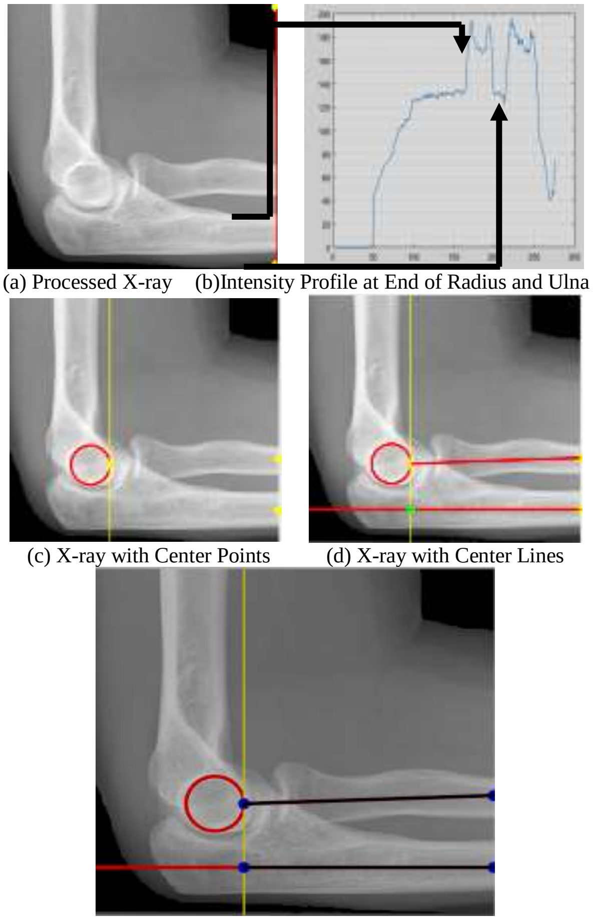
Fig. 5. Capitellum Detection on the X-ray Image of Elbow. (a) Indicates the Circular Shaped Capitellum Detection on Canny Mask. (b) Indicates the Capitellum Visualization on Actual X-ray Image.

2) Radial or ulnar fracture recognition: This test analyzes those elbow X-rays which clear capitellum based test and categorized as normal. For further diagnosing these two bones which runs from elbow to forearm, the center points of radius and ulna bones are identified by checking intensity at the end of these bones and using peak analysis as shown below in Fig. 6(b).

After getting the centerpoints as shown in Fig. 6(c) below, the centerlines are generetad from these points towards capitellum, these center points will serve as ending points to check intensity and the starting points are detected by placing a straight line at the maximum position of capitellum using basic parametric equation of line as shown in Fig. 6(d).

A number of points are located in between these points to check the intensity at the center of bones to detect any fracture in radius and ulna as shown in Fig. 6(e).

There are two possibilities of deformation in forearm bones as two bones are joint with capitellum.



(a) Processed X-ray (b)Intensity Profile at End of Radius and Ulna

(c) X-ray with Center Points (d) X-ray with Center Lines

(e) X-ray with Center Points for Intensity Check

Fig. 6. Processed X-ray for Radial or Ulnar Distortion Detection. (a) Shows the X-ray Image Processed Till the end of Radius and Ulna Bones to Check Intensity. (b) Shows the Intensity Profile at the End of Both of the Bones. (c) Shows the X-ray with Center Points Detection at the End of Ulna and Radius. (d) Shows the Processed X-ray with Center Lines in Middle of Forearm Bones. (e) Shows X-ray with Multiple Points at the Center Lines to Check any Fracture.

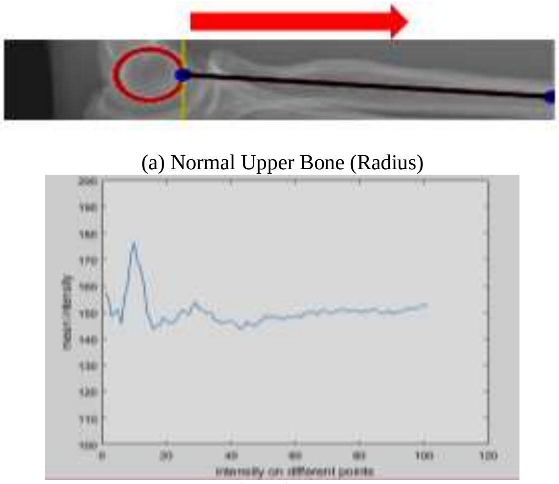
a) Normality detection in radius bone: The intensity is checked at each point in the center of normal radius bone as shown below in Fig. 7(a). The arrow points towards direction of intensity profile.

Due to shadow on the X-ray images at capitellum there is a little rise in intensity in Fig. 7(b). Except that, the intensity has same patter for normal radius bone.

b) Deformity detection in radius bone: The same procedure is applied on abnormal bone as shown in Fig. 8.

As moving far from the circular shaped capitellum, there is a fracture in the bone of radius which is detected in intensity profile as disturbance in intensity.

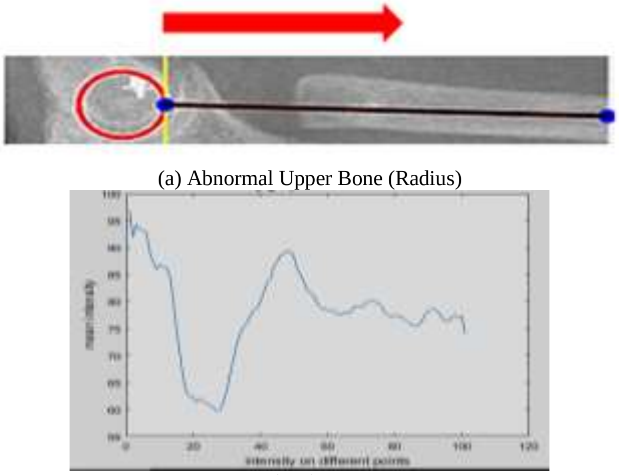
c) Normality detection in ulna bone: For normal lower bone (ulna), the same method is applied to check any deformation and the results are shown below in Fig. 9.



(a) Normal Upper Bone (Radius)

(b) Intensity Profile of Normal Upper Bone

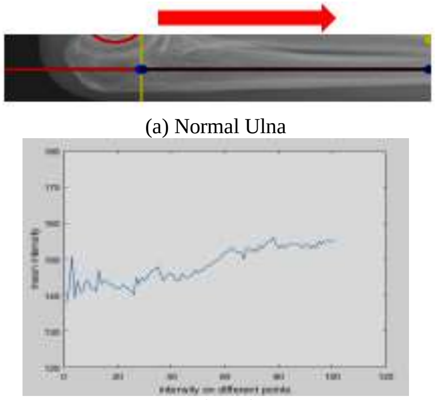
Fig. 7. Deformation Detection in Normal Upper Bone (Radius). (a) Shows the Usual Radius Bone with Multiple Points at Centerline to Check Intensity. (b) Demonstrates Intensity Profile of Usual Radius Bone.



(a) Abnormal Upper Bone (Radius)

(b) Intensity Profile for Abnormal Radius

Fig. 8. Deformation Detection in Abnormal Radius. (a) Indicates the Abnormal Radius Bone with Multiple Points at Centerline to Check Intensity. (b) Shows Intensity Profile for Abnormal Upper Bone (Radius).



(a) Normal Ulna

(b) Intensity Profile of Normal Ulna

Fig. 9. Deformation Detection in Normal Ulna. (a) Indicates the Normal Bone with Multiple Points at Centerline to Check Intensity. (b) Indicates Intensity Profile of Normal Ulna.

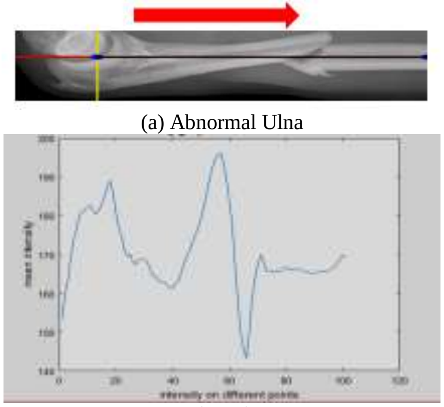
The intensity has almost same pattern for normal radius bone as there is no any fracture in the bone when checked by moving away from capitellum.

d) Deformity detection in ulna: The same process is practiced on fractured ulna and the intensity has a different pattern as compared with normal bone, the results are shown below in Fig. 10.

The intensity pattern for abnormal bone is totally different from the normal one as there is drop in intensity due to fracture in ulna bone as shown above in Fig. 10(b).

All the elbow X-rays having fracture in radius and ulna bone will be categorized as abnormal by this test and the ones categorized as normal will be further processed to check deformation, if any.

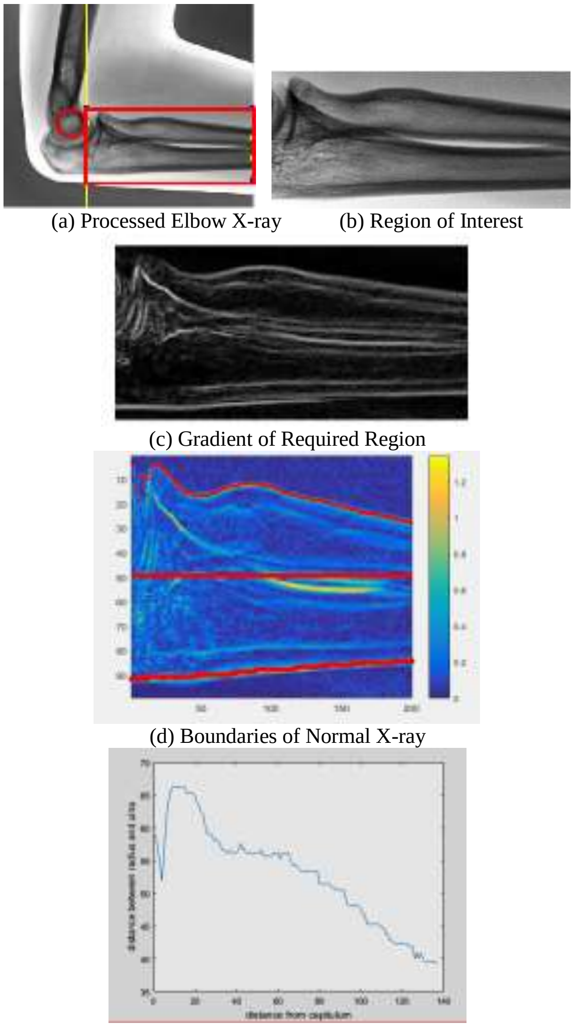
3) Geometrical shape abnormality identification: The last test will analyze those elbow X-rays which didn't have any visible fracture but still they are abnormal. This is possible when the bones are overlapped or displaced. To detect this type of abnormality, all the X-rays which intent capitellum are further processed by selecting the region of interest for this test as shown in Fig. 11.



(a) Abnormal Ulna

(b) Intensity Profile of Abnormal Ulna

Fig. 10. Deformation Detection in Abnormal Bone. (a) Indicates the Abnormal Ulna with Multiple Points at Centerline to Check Intensity. (b) Shows Intensity Profile of Abnormal Ulna Bone.



(a) Processed Elbow X-ray

(b) Region of Interest

(c) Gradient of Required Region

(d) Boundaries of Normal X-ray

(e) Distance between Boundaries of Normal Bones

Fig. 11. Geometrical Shape of Normal Bone. (a) Shows the Processed Elbow X-ray to get the Required Region. (b) Shows the Required Region to Detect the Abnormality in Bones. (c) Shows the Output of Image Gradient. (d) Shows the Colored Version of Gradient Output with Boundaries of Normal Bones. (e) Indicates the Distance between Normal Radius and Ulna Bone.

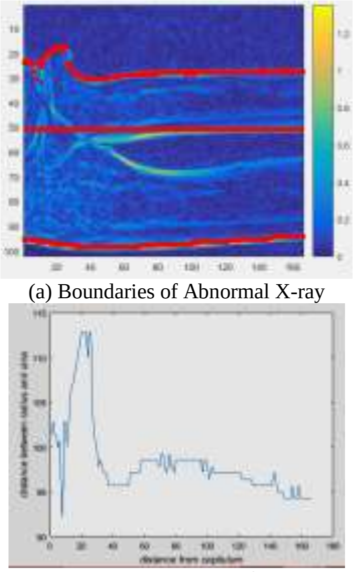
Image gradient [17] is applied on the selected region to measure the change of intensity as shown in Fig. 11(c).

For better visualization, color map is applied to the output of gradient and the region is split into two parts. Using peak analysis, local maxima is detected at every point to get the margin of bones as shown in Fig. 11(d).

Subsequently, the distance among normal upper (radius) and lower (ulna) bone is calculated and the significant pattern is examined in most of the normal images as shown in Fig. 11(e).

When the similar pattern is applied on abnormal X-ray with no visible deformation, the significant pattern analyzed in many normal X-rays isn't found which can be seen in Fig. 12(b).

Any X-ray which clears these three tests will be considered as normal elbow X-ray. A number of X-rays are analyzed and the cumulative results is shown in next section.



(a) Boundaries of Abnormal X-ray

(b) Distance between Boundaries of Abnormal Bones

Fig. 12. Geometrical Shape of Abnormal Bones. (a) Shows the Image with Boundaries of Abnormal Bones (Radius and Ulna). (b) Shows the Distance of Abnormal Bones.

IV. Result and Discussion

The three tests are suggested to automatically detect any deformation in elbow and connected bones. From the dataset, 250 elbow X-rays are processed for this research and has been checked by all three tests.

A. *Result for Capitellum based Test-01*

All of the 250 images are analyzed through this test to detect the capitellum as it is the major sign for normality and abnormality detection. All those X-rays which didn't reflect capitellum are categorized as abnormal X-rays and the ones that clears this test are further processed by other two tests to check any likelihood of deformation.

Amongst 250 processed elbow X-rays, 190 X-rays are standard (normal) and 60 X-rays are deformed (abnormal). The capitellum based test correctly interpreted 170 normal X-rays and 53 abnormal X-rays with the accuracy of 89.2% as shown below in Fig. 13(a).

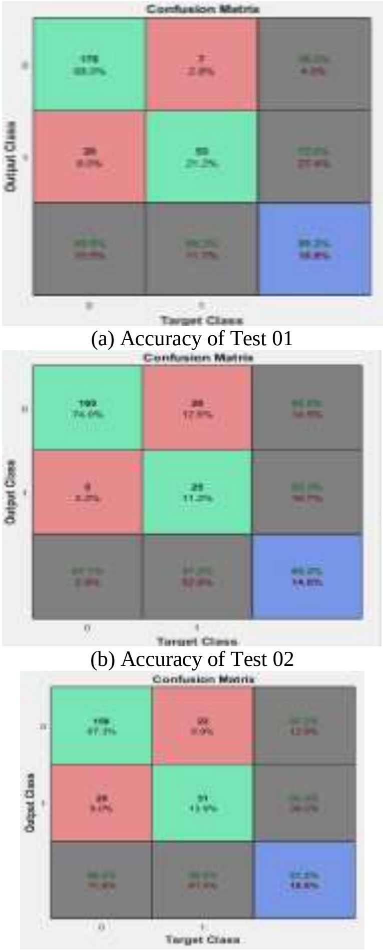
B. *Result for Radial and Ulnar Fracture Test-02*

All 223 X-rays that clears the previous test are further processed by this test. To detect fracture in radius and ulna bones, this test results in correctly analyzing 165 out of 170 normal and 25 out of 53 abnormal X-rays with the accuracy of 85.2% as shown below in Fig. 13(b).

C. *Result of Geometrical Shape based Test-03*

To detect any geometrical shape deformation, this test also analyzes those X-rays which cleared capitellum based test. 150 out of 170 X-rays are categorized as normal and 31 out of 53 as abnormal. This test results in accuracy of 81.2% as shown below in Fig. 13(c).

The overall accuracy of all the suggested tests for this study is observed as 86.2%.



(a) Accuracy of Test 01

(b) Accuracy of Test 02

(c) Accuracy of Test 03

Fig. 13. Accuracy of suggested Test to Automatically Detect Deformation. (a) Shows the Accuracy of Capitellum based Test 01. (b) Shows the Accuracy of Radial and Ulnar Fracture Detection Test 02. (c) Shows the Accuracy for Test 03 related with Geometrical Shape based Deformation Detection.

V. Conclusion and Future Work

In this study, a stimulating task to detect abnormalities of the bones in elbow X-ray images using computed approaches has been completed. 250 X-ray images from MURA dataset has been scrutinized for this process and the deformation in elbow X-ray images using computer based methods is detected centered on 3 Tests. The accuracy of proposed solution has been achieved around 86.2%.

Computer assisted techniques for finding deformation of bones in X-ray images will minimize human efforts and dependency which is the main problem in current time domain. Depending on those X-ray images which failed under these three tests, more features can be identified on the basis of edges and angular distance in between connected bones and the accuracy can be further improved.

References

[1] Kozaci N, AyMO, AkçimenM, Turhan G, SasmazMI, Turhan S, et al. Evaluation of the effectiveness of bedside point-of-care ultrasound in the diagnosis and management of distal radius fractures. Am J Emerg Med 2015, 33(1):67–71.

[2] J. C. McGinley, N. Roach, B. C. Hopgood, and S. H. Kozin, “Nondisplaced elbow fractures: A commonly occurring and difficult diagnosis,” Am. J. Emerg. Med., vol. 24, no. 5, pp. 560–566, 2006.

[3] P. Rajpurkar et al., “MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs,” no. Midl 2018, pp. 1–10, 2017.

[4] M. R. Zare, A. Mueen, W. C. Seng, and M. H. Awedh, “Combined feature extraction on medical x-ray images,” Proc. - 3rd Int. Conf. Comput. Intell. Commun. Syst. Networks, CICSyN 2011, pp. 264–268, 2011.

[5] American Cancer Society, Imaging (radiology) tests, 2013. Accessed on: August 2020. [Online]. Available: http://www.cancer.org/acs/groups/cid/documents/web content/003177

[6] S. M. Kengyelics, L. A. Treadgold, and A. G. Davies, “X-ray system simulation software tools for radiology and radiography education,” Comput. Biol. Med., vol. 93, no. July 2017, pp. 175–183, 2018.

[7] Y. Zhou et al., “Computer-Aided Detection (CADx) for Plastic Deformation Fractures in Pediatric Forearm,” Comput. Biol. Med., vol. 78, pp. 120–125, 2016.

[8] Suresh, S.S., 2009. Type 4 capitellum fractures: Diagnosis and treatment strategies. *Indian journal of orthopaedics*, *43*(3), p.286.

[9] Kurniawan, S.F., PUTRA, D., GEDE, I.K. and SUDANA, A.K.O., 2014. BONE FRACTURE DETECTION USING OPENCV. *Journal of Theoretical & Applied Information Technology*, *64*(1).

[10] Kaur, T. and Garg, A., 2016. Bone Fraction Detection using Image Segmentation. In*ternational Journal of Engineering Trends and Technology (IJETT)*, *36*(2), pp.82-87.

[11] Kim, H.T., Can, L.V., Ahn, T.Y. and Kim, I.H., 2017. Analysis of Radiographic Parameters of the Forearm in Traumatic Radial Head Dislocation. *Clinics in orthopedic surgery*, *9*(4), pp.521-528.

[12] Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., Ball, R.L. and Langlotz, C., 2017. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. arXiv preprint arXiv:1712.06957.

[13] Bishop, C.M., 2006. Pattern recognition and machine learning. Springer Science+ Business Media.

[14] Li, J. and Ding, S., 2011, August. A research on improved canny edge detection algorithm. In International Conference on Applied Informatics and Communication (pp. 102-108). Springer, Berlin, Heidelberg.

[15] Hawks, J. 2019. Humerus. [online] john hawks weblog. Available from: http://johnhawks.net/explainer/laboratory/humerus.html [Accessed 17th Jul. 2019].

[16] Djekoune, A.O., Messaoudi, K. and Belhocine, M., 2016. Incremental Hough transform: a new method for circle detection. In Computational Intelligence (pp. 3-22). Springer, Cham.

[17] Saif, J.A., Hammad, M.H. and Alqubati, I.A., 2016. Gradient based image edge detection. International Journal of Engineering and Technology, 8(3), p.153.

Compact Scrutiny of Current Video Tracking System and its Associated Standard Approaches

Karanam Sunil Kumar[1]
Assistant Professor
Department of Computer Science and Engineering
RNS Institute of Technology, Bangalore, India

Dr. N P Kavya[2]
Professor
Department of Computer Science and Engineering
RNS Institute of Technology, Bangalore, India

***Abstract*—With an increasing demands of video tracking systems with object detection over wide ranges of computer vision applications, it is necessary to understand the strengths and weaknesses of the present situation of approaches. However, there are various publications on different techniques in the visual tracking system associated with video surveillance application. It has been seen that there are prime classes of approaches that are only three, viz. point-based tracking, kernel-based tracking, and silhouette-based tracking. Therefore, this paper contributes to studying the literature published in the last decade to highlight the techniques obtained and brief the tracking performance yields. The paper also highlights the present research trend towards these three core approaches and significantly highlights the open-end research issues in these regards. The prime aim of this paper is to study all the prominent approaches of video tracking system which has been evolved till date in various literatures. The idea is to understand the strength and weakness associated with the standard approach so that new approaches could be effectively act as a guideline for constructing a new upcoming model. The prominent challenge in reviewing the existing approaches are that all the approaches are targeted towards achieving accuracy, whereas there are various other connected problems with internal process which has not been considered for e.g. feature extraction, processing time, dimensional problems, non-inclusion of contextual factor, which has been an outcome of the proposed review findings. The paper concluded by highlighting this as research gap acting as contribution of this review work and further states that there are some good possibilities of new work evolution if these issues are considered prior to developing any video tracking system. Overall, this paper offers an unbiased picture of the current state of video tracking approaches to be considered for developing any upcoming model.**

Keywords*—*Video tracking; object tracking; visual tracking; video surveillance; object detection

I. Introduction

With the advancement of computer vision and video surveillance systems, video tracking has gained immense popularity in both domestic and commercial applications [1]. Fundamentally, video tracking is a mechanism of identifying, recognizing, and tracking a mobile object over time [2]. Apart from its applicability towards video surveillance systems, video tracking is now used over various applications: viz. video editing, medical imaging, traffic control, augmented reality, communication, and video compression, human and computer communication [3-5]. Usually, the comprehensive mechanism of a video tracking system could involve more processing of time owing to its dependency on a massive amount of data within a video sequence [6]. Complexity in the operational process also existing in recognizing an object with accuracy in a video tracking system [7]. Essentially, the video tracking system aims to connect the mobile target object (or multiple objects) present over a sequence of video frames. This could be highly a difficult process, especially when the speed of the mobile object is quite faster relatively or uncertain concerning the defined rate of video frames. The uneven orientation of a mobile object is another complicated scenario in video tracking, which offers complexity in analyzing the presence of an object for a given scene over a sample of time. In order to address this conventional issue, the motion model is adopted in the video tracking system [8]. This motion model is responsible for defining the relationship between the target object image and its influence over the mobility scenario. Regarding the motion model, generally homography or affine transformation is used for two-dimensional models when tracking is performed over planer objects [9]. The motion model for a three-dimensional object is usually related to the position and orientation of the object [10]. While dealing with video compression, the macroblocks are divided into keyframes, and selected motion motions are considered disruptions of these frames considering motion parameters [11]. In the case of a deformable object, the motion model generally considers the position of a target object over the mesh [12]. At present, there is various literature on video tracking systems, which mainly evaluates sequential frames in a video yielding to an identified target object within the transition of frames [13-16]. However, considering the generalized classification, it is found that existing video tracking algorithms are of two types, i.e., representation along with localization of a target object and filtering of data. The first kind of algorithm are generally known for their low computational complexity, and they are again classified into contour-based tracking and kernel-based tracking. The second kind of algorithm mainly deals with the dynamics of the target object and performs assessment based on multiple hypotheses. Thereby, such an algorithm results in enhance capability towards tracking mobile objects of complex form. However, these algorithms are also computationally complex, and it has dependencies over different parameters, e.g., stability, redundancy, quality, etc. The algorithms that fall under such category are Kalman filter and particle filtering. Therefore, the prime research problem considered for this work is that – although there are various implementation and discussion-based papers on video tracking system, but there is no global-viewpoint to standardize the effective of all the exercised approaches. It is still vague to understand the actual scenario of existing video tracking approaches as the taxonomies are not well discussed. Apart from this, it is also known fact that there is an increasing demands of video surveillance system where sophisticated features are demands. There are also consistent evolution of various approaches in order to assist in internal processing of video frames in image processing. This leads to a motivating factor that this topic is worth doing a research on owing to its abundant scope of application in upcoming days as well as trade-off in finding any potential standardized model.

Therefore, the significant objective / contribution of this paper is to discuss the techniques applied in major implementation work towards the significant classes of video tracking approaches, i.e., point-based tracking, kernel-based tracking, and silhouette-based tracking. The study also contributes to discuss open research problems. The organization of this manuscript is as follows: Section II briefs about the taxonomy of video tracking approaches followed by a discussion of Point tracking approaches in Section III. Existing kernel tracking and silhouette tracking approaches are carried out in Section IV, and Section V. Discussion of findings of the study is carried out in Section VI. A summary of this paper is carried out in Section VII.

II. Taxonomy of Video Tracking Approaches

Basically, the video tracking mechanism targets trajectory generation associated with an object by identifying the position of an object with respect to all the video sequences over time. The taxonomy of the existing video tracking approaches is pictorially shown in Fig. 1. It is found that existing video tracking approaches are broadly classified into three forms, i.e., point-based, kernel-based, and silhouette-based approach. The point-based approach makes use of points for tracking and is further classified into deterministic and probabilistic approaches. In a deterministic approach, the cost factor is evaluated for connecting with each object over a video sequence considering motion constraints (i.e., common motion, change of minor velocity, higher velocity, rigidity, and proximity). In the deterministic method, the state-space approach is usually used to model the properties of an object and its associated parameters (i.e., acceleration, velocity, position, etc.). At present, studies towards point-based tracking mainly use Kalman filtering, particle filtering, and hypothesis-based tracking system. The kernel-tracking-based system uses the mobility of an object with respect to connecting frames and is broadly classified into two forms, i.e., multiple tracking and template-based tracking. The multi-view tracking system is modeling without the inclusion of zero interactivity between object and background. The aggregated information is represented by this model, corresponding to the objects in the given scene of the video sequence. The possibility of different shapes and sizes of the same object is quite high in this approach. It is further categorized into subspace and classifier forms of tracking. Template-based matching is used mainly for tracking single objects, and it ensures the minimal cost of computation. It has been found that the kernel-based tracking dominantly uses the mean-shift approach, support vector machine, and template-based approach. Such an approach is also discussed to offer better occlusion handling capability while requiring more training to attain better performance. The final class of video tracking system is silhouette tracking mechanism, which is meant for overcoming the incapability in tracking issues associated with simplified geometric shapes, e.g., shoulders, head, hands, etc. Therefore, in this form of tracking, the region of the objects is explored from each frame to offer a detailed description of the target object. The modeling is feasible in this form of tracking mechanism using contours of an object or edges of an object or color histogram.

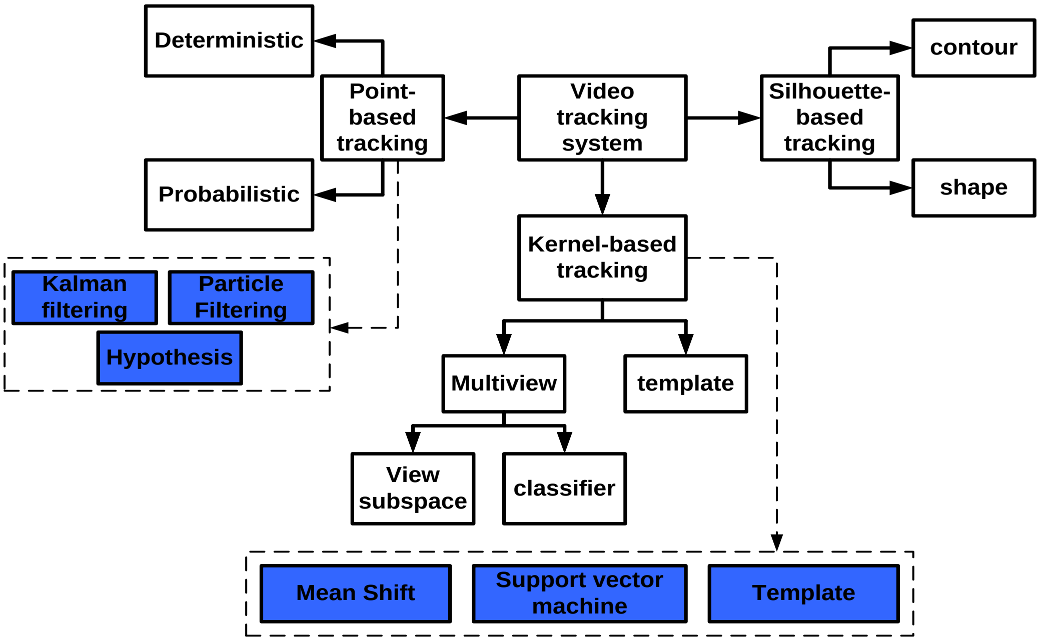


Fig. 1. Taxonomy of Existing Video Tracking System.

Fundamentally, there are two classes of silhouette tracking systems, i.e., shape-based and contour-based. The shape is also used to define the state-space of an object. In order to increase the posteriori probability, the system always updates the state over a specific period of time. In such a case, the posterior probability depends upon the initial state and the likelihood of the current state that generally uses the spatial distance between two contours. It is to be noted that the silhouette-based approach classification is more or less a similar form, i.e., contour and shape-based, and no new forms have been ever surfaced. In all the approaches mentioned above, certain issues have been consistently under observation, i.e., occlusion and tracking using multiple cameras. There are three classes of occlusion, i.e. i) Occlusion due to structures in the background scene, ii) Inter-object occlusion, and iii) Self-occlusion. Developing a uniform algorithm for video tracking, considering all these three occlusion cases itself, is a challenge of a bigger dimension. Similarly, variation in shape and size of the same object is a bigger challenge when applying a multi-view tracking system. The next section briefs about existing research work in this direction.

III. Point Tracking Mechanism

Point tracking is the first form of the object tracking method, where various forms of points are used across the target frame to represent the identified object over the video sequences. However, mapping a specific point over an identified object is quite a complex scenario, especially when a target object exists in the scene, misdetection, or occurrence of occlusion. Basically, this system is of two types, i.e. deterministic approach and probabilistic approach. However, a closer look into the trends of the approaches and methods being carried out by point-based tracking system is mainly found to use Kalman filter, particle filtering, and hypothesis-based tracking.

A. Kalman Filter-based Methods

In recent years, the Kalman filter usage has significantly increased for object detection under various environmental conditions. In most of the work, the adoption of the Kalman filter is proven to offer significant accuracy when it comes to tracking at a higher speed. Even under the complex form of video files (e.g., satellite videos), the Kalman filter is reported to offer better tracking performance (Guo et al. [17]). Global motion attributes characterize the moving object to offer a measurable score of tracking performance using the Kalman filter. The core part of the tracking system is developed using the correlation filter, which uses the original pixel and HOG to represent its features. The study formulates an objective function intending to reduce the squared form of the error that occurs between the suitable map of response and correlated response. The study integrates the usage of correlation filter and Kalman filter to facilitate higher tracking speed with accuracy and fault tolerance. However, the limitation is that usage of this approach requires more robustness while performing dynamic tracking operations. A study towards achieving robustness is carried out by Gupta et al. [18], where the depth of interest is used to perform object tracking over the mobile environment. The study makes use of an unscented Kalman filter using an experimental approach for performing forecasting of the location of a mobile object. A similar direction of the work towards tracking a mobile object's position is also discussed by del Rincon et al. [19]. In this work, the author has considered a use case of tracking different parts of the human body using the strategy of multiple tracking and a two-dimensional articulated model. The interesting part of this study is its supportability towards identifying and tracking various rotational aspects of the human body. Study towards tracking multiple targets have been carried out by Wang and Nguang [20]. The uniqueness of this part of the study is to integrate the connected data using a probabilistic model with the Kalman filtering (Fig. 2).

A slightly unconventional mechanism of object tracking is carried out by Yang et al. [21] considering the use case of tracking aircraft. The study has used a deep learning mechanism to improvise the accuracy in the tracking system where the model integrates the Kalman filter and extended Kalman filter to forecast the trajectory. Based on a region-based deep neural network, the presented scheme uses a shared structure of the convolution, which is used to encode the data connected with the positional information of the flying object. The region of interest area is then attached to the pooling layers on the top of the deep neural network where the response system of (i, j) is mathematically depicted as:

$$r_c(i,j|\theta) = \sum_{(x,y)\epsilon bin(i,j)} z_{i,j,c}(x + x_0, y + y_0|\theta) \quad (1)$$

In the expression (1), the variable r_c represents the mapping score associated with the region of interest. Each class is further subjected to a software response system in deep learning. The mobility model of the presented deep learning mechanism is as follows:

$$x_k = \mathbf{A}x_{k-1} + Bu_k + w_k z_k = \mathbf{H}x_k + v_k \quad (2)$$

The above expression (2) is used for tracking the mobile target where state vector x is used to define the system with estimated value as z and matrix for state transition $\mathbf{A}$. The variables B and u represent the controlling parameters, while the transformation matrix is represented as $\mathbf{H}$. This model also considers the presents of noises w and v. The study outcome (Fig. 3) shows that the model is capable of identifying the flying object under different context of background. Irrespective of any direction of mobile state of air-born object, the model can successfully perform identification. The study outcome has been finally verified by comparing with other existing contents on multiple schemes, e.g., mean strategy, cropping with correction, cropping with estimation, normal cropping, and estimation.

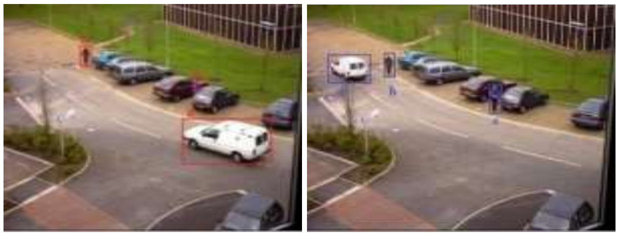


Fig. 2. Visual Outcomes for Multi-Object Tracking Wang and Nguang [20].

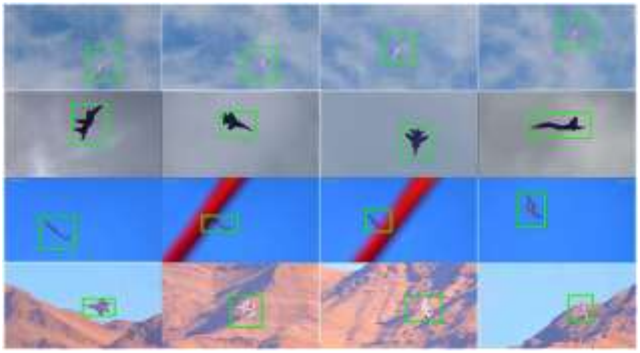


Fig. 3. Visual Outcomes of Tracking Yang et al. [21].

Table I highlights that Kalman filtering with deep learning, offering the higher capability to perform tracking of the dynamic mobile object. The study has also claimed to reduce the detection time; however, the spontaneity in the tracking duration may differ based on the different background aspects of the scene, which is not found to be discussed in the presented system and thereby acts as limiting factor.

B. Particle Filtering Method

When there is a large set of information, the data sample is required for performing any further processing. This sample of data, which is also called a particle, is utilized to represent the data distribution associated with various stochastic nature processes. The particle filtering process is used for extracting such filtered samples of a particle in the presence of noisy information. There is also a higher possibility of the presence of impartial information and a non-linear state of varied form. From the perspective of the video tracking system, there is a need to track the object with various nonlinearities and uncertainties. Hence, the concept of particle filtering suits well in designing a video tracking system. One of the significant advantages of using the particle filtering process is its inclusion of a non-Gaussian mechanism of distribution study and nonlinearities. Apart from this, it can also be said that particle filter acts as a better alternative for the existing Kalman filter. The critical issue associated with Kalman filtering is that it assumes a normal distribution of state variables, which is less practical in nature and therefore it is its limitation. Such issues can be addressed using particle filtering where the state density at a specific time can be mathematically represented as:

$$\text{sample}_n t : n=1, \ldots., N \qquad (3)$$

The variable *n* represents particles with a sampling probability of π_t^n as weight, which can index the significance of the considered sample. This mechanism can also address the computational complexity by storing the cumulative weight for each tuple. The frequently used sampling process includes selection, prediction, and correction. In the *Selection* method, the random samples s_t^n is selected from S_{t-1} by generating an arbitrary number *r* in the probability range of [0, 1]. The idea is to find the smallest sample j such that sampling probability c_{t-1} is less than *r* considering $s_t^n=s_{t-1}$. In the *prediction* method, a new sample is generated as $s_t^n=f\{s_t^n, W_t^n\}$, where *f*(x) is a non-linear function with W_t^n as mean Gaussian error. In the *correction* method, the estimation of weights $\pi_t^n$is carried out where π_t^n is equivalent to $g(z_t|x_t = s_t^n)$, where g is Gaussian density function. Adoption of these methods offers more comprehensive tracking performance, even considering a different number of features. One such recent work used a similar approach where multiple features are used towards facilitating video tracking (Bhat et al. [22]). The authors commented that features are exclusive of environmental variables, and various attributes of color distributed can be used as feature space. It is highly application-specific, while the study has considered that the KAZE feature, which is capable of blurring and smoothening the information along with noise. This challenge is addressed by using additive operator splitting for achieving the sharpness. According to this model (Fig. 4), the system takes the input of video sequences followed by selecting a target from a specific frame. The particles are generated in the surrounding of the centroid of the blob, followed by updating the particles. This updating procedure can be carried out by using the motion model. Finally, the particles based on the spatial score are weighted, followed by resampling all such particles to obtain a new centroid, which leads to the generation of the filtered location of the target, thereby assisting in video tracking.

TABLE I. SUMMARY OF OUTCOME YANG ET AL. [21]

Approach	Accuracy(d>0.8)	Accuracy(d>0.9)
Yang et al. [21]	**95.9**	**95.2**
Mean strategy	94.3	93.0
Crop, correction	92.3	91.6
Crop, estimation	90.5	89.2
Normal crop	87.9	86.1
Estimation	88.5	86.1

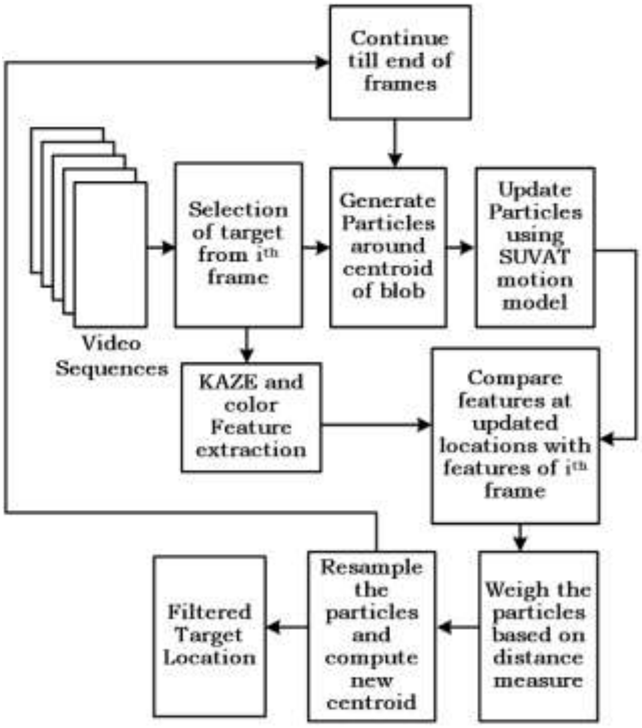

Continue till end of frames
Video Sequences
Selection of target from i^th frame
Generate Particles around centroid of blob
Update Particles using SUVAT motion model
KAZE and color Feature extraction
Compare features at updated locations with features of i^th frame
Filtered Target Location
Resample the particles and compute new centroid
Weigh the particles based on distance measure

Fig. 4. Usage of Particle Filtering (Bhat et al. [22]).

Particle filtering is also used to address the issues of the appearance model, which suffers from various extrinsic limitation factors, e.g., clutter in the background, occlusion, and variation in illumination (Wang et al. [23]) (Fig. 5). This system uses unique particle filtering to generate information about the state of the target concerning the current frame immediately after updating the template.

It should be noted that the study mentioned above is based on tracking using a matching mechanism also, which has a dependency upon an interesting local point, thereby reducing the robustness. This problem is sorted out by Zhang et al. [24], where basis matching is used as a substitution of point matching. Gabor filter is used to learn the target model, while particle filter identifies the object over the dynamic system. The study outcome was assessed on various test environments of occlusion, variation in illumination condition, alteration in poses of an object, and clutter in the background.

C. *Hypothesis based Method*

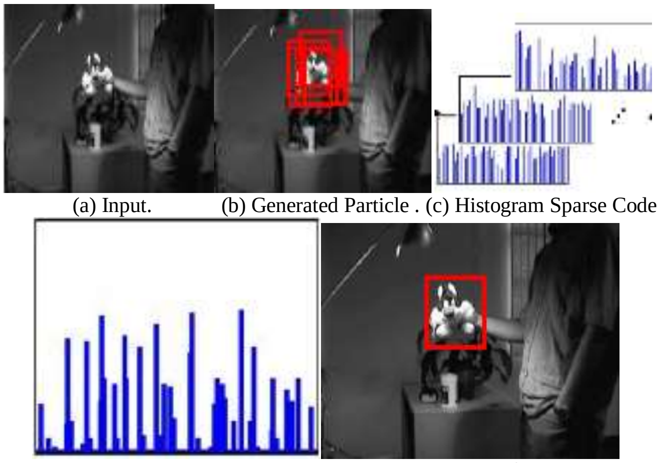
The majority of the video tracking system consists of the inclusion of two video frames, where there is less likelihood of inappropriate correspondence if the correspondence is created between two frames only. It facilitates effective tracking outcomes when there is a deferment of reading other video frames. In this regard, this process of multiple hypotheses helps manage multiple such correspondence hypotheses associated with all the objects for the given instantaneous frame. This approach offers a higher likelihood of the last frame with an object over a specific time period with a capability to construct upcoming queued tracks for the next object while eliminating the already existing track results. It should be noted that multiple hypothesis-based approaches are essentially an iterative process that initiates from the set of current tracks while multiple disjoint tracks formulate to form a collection within the hypothesis. The system then carries out a prediction for the position of an object for each hypothesis over the consecutive frame. These predictive outcomes are compared with the original measurement by assessing spatial measurement. Depending on this measurement of the spatial score, the system establishes hypotheses that further provide new hypotheses over the next rounds of iteration. However, it is necessary to know that owing to the iterative operation. It leads to a computational burden. This complexity can be addressed by considering probabilistic modeling, where the correspondence is random variables that are statistically independent of each other. Particle filtering can also be used to address this issue; however, it may offer a lower probability of enumerating all the possible correspondence. Hence, multiple hypotheses area better option when it comes to the demand for checking all the possibilities.

Another advantageous feature of the multiple hypothesis-based tracking systems is their ability to perform tracking of smaller targets. However, it is associated with the larger tree structure in existing approaches with a large number of branches. This issue can be sorted out by applying a certain optimization approach. Work towards this target is carried out by Ahmadi and Salari [25], where particle swarm optimization has been used to explore the optimal number of tracks from the video sequence. The implemented steps in this work are i) exploring the preliminary tracks with the aid of a multiple hypothesis approach, ii) fine-tune and adjust the observed track information using particle swarm optimization, and iii) merging all the collected track information that maps with a single target object. However, the limitation of this approach is its capability to track only a single object.

This limitation is overcome in the work of Kutschbach et al. [26], which extends its tracking towards multiple objects using the probability of Gaussian mixture with multiple hypotheses. The study also makes use of a kernelized correlation filter for better accuracy performance. It is to be noted that the iterative nature of this approach is also discussed in existing literature for optimal outcomes. However, most of the existing approaches are found to have a lack of any inclusion of relevance between two video frames which is one major limitation. Apart from this, there is no optimized approach to utilize the preliminary information from the individual frames (Sheng et al. [27][28]). The optimization carried out in this approach is to undertake the information about independent sets of maximum weight. The study constructs the hypothesis between the consecutive video frames using the transfer model of the hypothesis. Also, the complexity associated with the iterative process has been addressed using an approximation algorithm of the polynomial time. This process indirectly improves the efficiency of the system. The upper bound *UB* of this tracklet is mathematically given in this work as.

$$UB(v) = \max\{w(F[v] - w(n^*), w(v) + w(n^*) + w(F(v) \cap F(n^*)\} \quad (4)$$

In the above expression (4), where v is a set of tracklet. A closer look at the above expression shows that there are two variables, i.e., F(v) and F[v]. The first variable F(v) represents the possible set of hypotheses for v tracklet, whereas F[v] represents union operation of F(v) with all the elements of the set of tracklet {v}. The variable w represents the weight function that assists in further offering informative significance about the identified object.



(a) Input. (b) Generated Particle . (c) Histogram Sparse Code.

(d) Observation Histogram. (e) Localization of Target.

Fig. 5. Visual Outcomes of Particle Filtering (Wang et al. [23]).

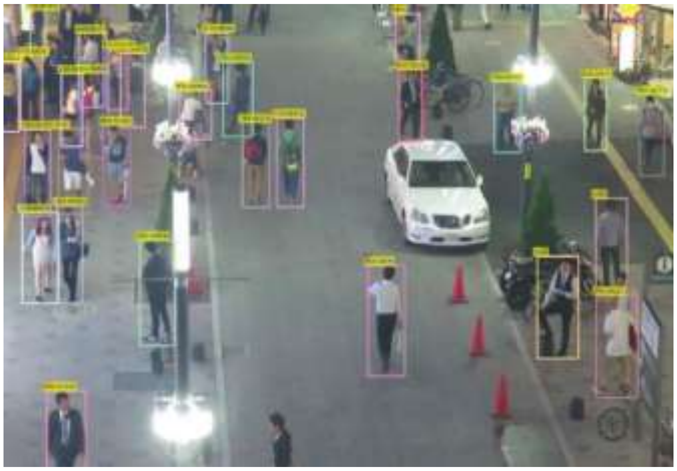


Fig. 6. Live Tracking over the different Size of the Window(Sheng et al. [27]).

Fig. 6 highlights the visual outcomes to show that this model can track different objects at the same time with different sizes of windows. However, this approach is limited to a single camera with multiple object tracking. Further, the authors have developed a graph model with distance and time information connected with the trajectories. The model has used a temporal graph to assess the presented tracklet generation, resulting in connectivity among hypotheses and benchmarking. The video tracking operation is further improved when this model is integrated with tracking using network flow. At the same time, similar network flow parameters are utilized to assess the validity of the model. The test environment used in this study is further extended where multiple similar targets are subjected to tracking but using multiple cameras (Yoo et al. [29]). The tracking is carried out over multiple tracks that are completely unknown and obtained from time and distance relationships. The realization of the multiple tracks is carried out by solving the clique problem of a higher degree of weights associated with each frame. The study makes use of feedback information obtained from the result of the preliminary frame online. With the adoption of the tracklet, the presented approach is now capable of generating much fine-tune set of candidate tracks and filtering out all the candidate tracks that are found to be unreliable. Hence, there are various point tracking systems at present for video tracking system.

IV. Kernel-based Tracking Mechanism

This is a typical mobile object tracking system in a video represented using a primitive object region from one to another video sequence frame. Normally, the parametric motion is witnessed in the form of affine, conformal, and translation for all the motion of objects. The computation of the flow field of dense nature can also be used to represent the motion of an object. Various approaches in such methodology are constructed based on the techniques used for motion estimation of an object and the number of a tracked object. The existing literature is witnessed to adopt mainly three essential approaches under this method viz. i) Mean Shift Method, ii) Support Vector Machine, and iii) Template-based Method.

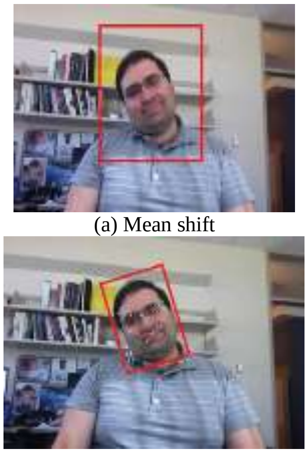
A. Mean Shift-based Method

Although this is one form of video tracking mechanism, its core principle is based on the video sequences segmentation. Existing literature discusses a technique where the mean shift approach is used along with the other associated techniques to improve the tracking performance. The work carried out by Baheti et al. [30] has used an enhanced Lucas Kanade Algorithm for effective controlling of the computational complexity for performing object tracking. The objective function stated for this purpose is:

$$error = \sum_{x}[I(W(x:p+\Delta p)) - T(x)]^{2} \tag{5}$$

In the above expression (5), the estimation of an error for aligning the template with reference is carried out by considering input image I with reference image T(x). W is considered as a warping function with warping parameters p such that $p=(p_1, p_2, \ldots, p_n)^T$, while Δpisincrement parameter of warping function. The preliminary set of warping parameters is obtained from the RANSAC algorithm using its homography estimation. The warped image is subtracted from the gradient, followed by computing the gradient information about the template image in a specific direction and extracting Jacobian related to it. The descent image of the steepest form computed using matrix multiplication followed by computing the Hessian matrix with further multiplying it with the error. The variation is the computed parameter, which is subjected to the objective function for minimization of the error leading to good accuracy in tracking.

Fig. 7 highlights the visual performance of both the method to show that adding the Lucas Kanade Algorithm with mean shift offers more accuracy compared to conventional mean shift. However, it should be noted that this approach doesn't emphasize much on the complex environment of background, which is necessary for adaptive tracking. The existing approach reports that the usage of the kernel correlation filter can solve this issue of complexity associated with the background when used alongside with mean shift (Feng et al. [31]). In this method, the trained image with its respective positional information is considered an input followed by tracking based on the kernel correlation filter.



(a) Mean shift

(b) Lucas Kanade

Fig. 7. Visual Outcome of Tracking Baheti et al. [30].

Further, with the inclusion of the new frame, the confidence map of the filter is obtained, which is followed by assessing if the mean shift is required to be used. For this purpose, the histogram feature is mean shift is obtained, which finally leads to the outcome of tracking (Fig. 8). However, the method doesn't include occlusion mitigation, which is required for cluttered scene analysis. The problem of occlusion and complex background have better possibilities of solving if the emphasis is offered much on data distribution with a multi-dimensional approach. The conventional mean-shift approach can be extended to three dimensional with more suitability in tracking dynamic object (Liu et al. [32]). Such mechanism performs dual steps viz. i) all the significant mobility points are tracked, and appearance model is subjected to related fine-tuning necessary and ii) detection process is initiated along with compensation of errors in tracking owing to complete occlusion. The study outcomes show some robust tracking performance compared to some of the existing mechanisms of different variants of kernel correlation filters using colored videos. The technique involves preprocessing the infrared sequence of video followed by target identification. For the first image, the detected target region is captured, followed by multiscale transform and fusion of the target region. Upon subjecting to transformation using multiscale image give the outcome for part of the fused image. A similar sequence of processes is carried out for the second image. The background is captured from the identified target, followed by a similar set of processing carried out by the first image to give a second set of fused images. Both the fused image is further organized as a sequence to perform tracking. A recent work carried out by Peng and Zhang [33] has a unique implementation of mean shift where detection and tracking of the target region are tracked using the mean shift method. In contrast, the root means square is estimated between two frames to assess the error score. Other associated studies on similar technique with slighted variation in using the mean shift was seen in the work of Shu et al. [34], Tan et al. [35], Wang et al. [36], and Chen et al. [37].

B. Support Vector Machine based Approach

In the area of the learning algorithm, Support Vector Machine (SVM) is considered as a supervised model which is capable of performing both linear and non-linear classification. This characteristic makes it suitable for improving the accuracy of the video tracking system. The SVM approach has proven effective when it comes to object recognition and tracking. SVM, when combined with Scale Invariant Feature Transformation (SIFT), offers better performance (Dardas and Georganas [38]). The technique applies vector quantization for mapping key points with the training image, followed by applying k-means clustering and bag-of-words. However, better classification performance is seen when one-class SVM is used with Markov chain Monte Carlo implementation (Feng et al. [39]). The inclusion of dynamics in tracking are addressed using probability hypothesis density. The enhancement of SVM in tracking is further proven when excluding the coupled label prediction using kernelized supported SVM for adaptive tracking. The complexity owing to unbound growth in support vector is controlled using a budgeting mechanism. This fact was also verified by Yuan et al. [40] using multiplicative kernels. Such approaches also void the inclusion of contextual modeling, which otherwise is discussed to offer better SVM predictability (Sun et al. [41]). Such an approach decomposes the spatial context in the form of foreground and background for obtaining a robust appearance model to deal with deformation and occlusion issue in video tracking. Sun et al. [42] have also used SVM for categorizing the scene from its sophisticated surroundings. Such an approach is proven to encode the perception of human vision using gaze shifting path.

Fig. 9 highlights the process used where an aggregated convolution neural network is used along with over gaze shifting path further subjected to SVM for effective classification. Such idea of the combined process, i.e., identifying an object, learning, and tracking, is also carried out by Yin et al. [43]. In this work, SVM is used for dual purpose, i.e., performing linear classification and state-based structure classification where applicability increases over complex video scenes. SVM is also proven to reliable modeling (Sun et al. [41][42]) where learning is carried out over multiple views and harnesses the geometrical structures of the tracked outcome. Overall, SVM has a fruitful performance when it comes to video tracking from complex video sequences. However, the approaches don't offer many solutions towards computational complexity associated with classification performance.

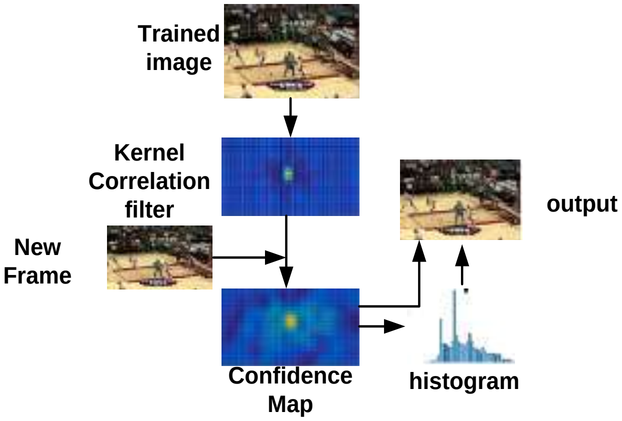


Fig. 8. Process Flow of Kernel Correlation Filter with the main Shift (Feng et al. [31]).

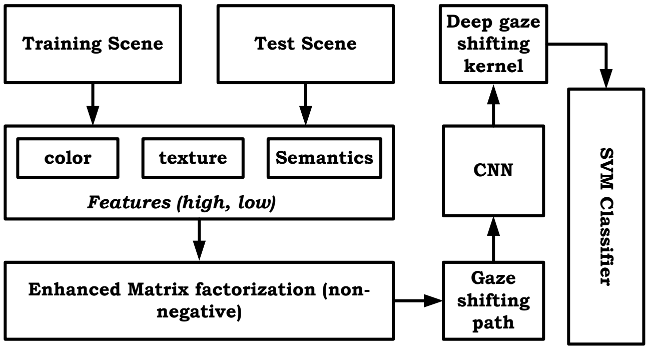


Fig. 9. Categorization of the Scene for Tracking Sun et al. [42].

C. Template-based Approach

The formation of the template is usually carried out using normal geometrical structures. It is capable of bearing both information of appearance and spatial data from the given scene. However, one of the pitfalls of this approach is that the generated appearance of an object can only be encoded from a single view. This narrows down its applicability towards tracking video with lesser variation in poses of an object while tracking. Hence, there are various attempts in present times to circumvent this issue. Guo et al. [44] have used an adversarial network with the guidance of the generative task to perform dynamic learning. Templates are selected from online adaption from the image sources with ground truth along with an arbitrary vector. However, it doesn't perform the dynamic matching of the template. This problem is discussed to be solved by Huang et al. [45] where segmentation of an object is carried out using aggregation network with temporal attribute with Hungarian matching scheme from template bank (Fig. 10).

Existing studies have also witnessed the template matching process to be hybridized where different methods from other categories of video tracking are found to be used. Studies of Lin and Chen [46] and Mutsam and Pernkopf [47] have discussed the usage of the particle filter with template matching. A unique study carried out by Pei et al. [48] has used graph matching over the template to establish the connection between objects and trajectories. Rehman et al. [49] have used template matching and deep learning over multiple regions of interest to improve its scope of tracking and accuracy. The work carried out by Su et al. [50] has integrated color histogram with a template for maximizing the performance of video tracking along with the significance of the update process over selected regions of interest. Xiu et al. [51] have extracted differential information where the initial region of the target is carried out using rough matching followed by magnifying the region of search. The study outcome is proven to offer better tracking performance in contrast to the existing template matching algorithm. Apart from this, various artifacts associated with the interference are also solved in such an approach with a reduction in outliers for improving video tracking performance.

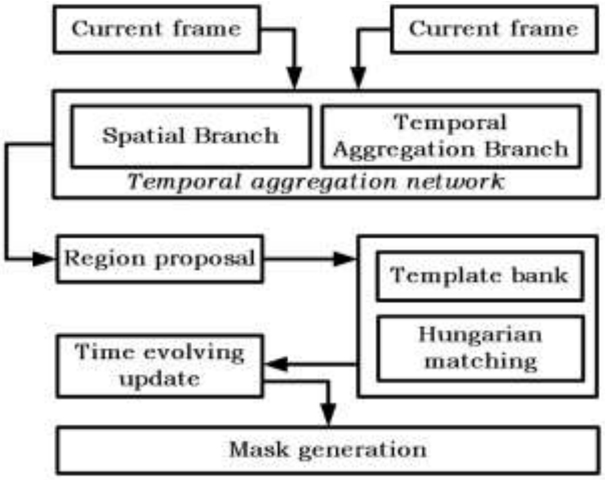


Fig. 10. Process Flow of Template Matching Huang et al. [45].

V. Silhouette Tracking Mechanism

The silhouette is a simpler mechanism for tracking an object when it comes to the non-rigid nature of an object. In such a case, the region of an object is estimated for all the frames to facilitate tracking. The encoded information with the region of an object is used for this form of the tracking system. The possibilities of such information could be an edge map, or it could also be any shape model. Typically, contours and shape factors are used in the process of a silhouette tracking mechanism.

Object tracking could be a complicated process when the video is of multi-dimension as there is a proliferation of multi-dimensional video with the advancement of digital technologies. The work carried out by Kim et al. [52] has addressed this problem by performing contour tracking using the graph-cut method. The study considers the distance of the angular radial factor and its variation as its essential constraint with a presence of deformation in shape. With the refinement of the contours, the ambiguous seeds are eliminated for precise segmentation using graph cut. Combined with a neural network, the performance of the contour-based tracking system can be enhanced (Kishore et al. [53]). This strategy uses the Horn Schunk optical flow method for obtaining features for tracking while the shape features are extracted from active contours. Different events are classified using the backpropagation approach in the form of words, and then converted into signals for matching. The work of Luo et al. [54] has implemented a silhouette-based tracking system using a segmentation approach with a block-based technique. The information of motion during the encoding of the video is utilized for tracking purposes. Kalman filter is also reported to enhance this in the form of video tracking system (Pokheriya and Pradhan [55]). The study makes use of the background subtraction method of adaptive nature.

Another unique mechanism to carry out this silhouette tracking mechanism by using the camshaft algorithm (Zou et al. [56]). This is mainly utilized for computing the distribution of color probability, thereby facilitating a video tracking system. Apart from this, the inference system becomes quite simpler concerning the background. This algorithm is considered to be useful to deal with occlusion and target deformation. Fig. 11 showcase the unique outcome where the original image (Fig. 11(a)) is processed to the obtained distribution of color probability (Fig. 11(b)) followed by extraction of distribution of motion probability (Fig. 11(c)) and distribution of cumulative probability (Fig. 11(d)).

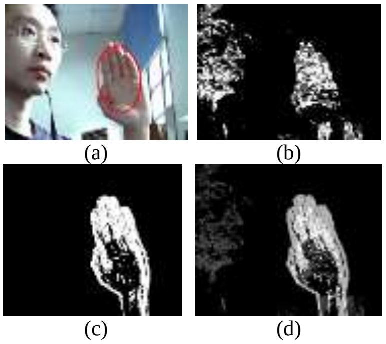


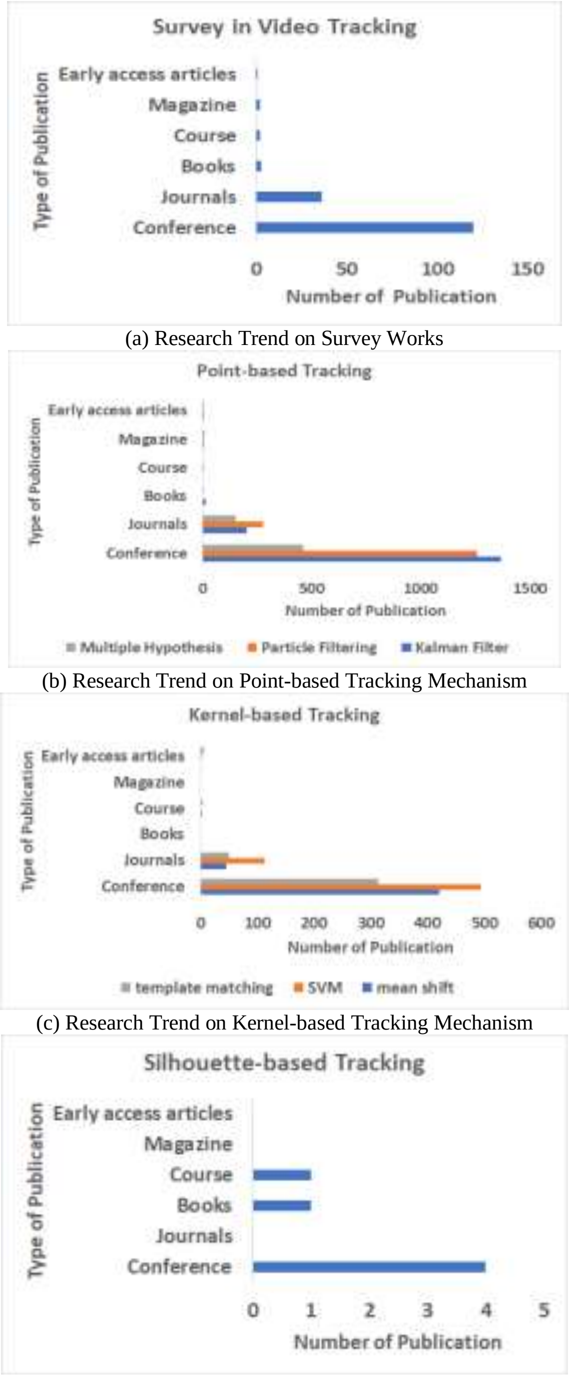
Fig. 11. Silhouette Tracking by Probability Distribution (Zou et al. [56]).

VI. Study Findings

The core study findings of the proposed paper are discussed with respect to existing research trends and briefing of open-end research problems.

A. *Existing Research Trend*

To visualize existing research trends, the proposed system collects the paper published in the IEEE Xplore digital library published between 2010 and 2020. The findings are graphically shown in Fig. 12 as following,



(a) Research Trend on Survey Works

(b) Research Trend on Point-based Tracking Mechanism

(c) Research Trend on Kernel-based Tracking Mechanism

(d) Research Trend on Silhouette-based Tracking Mechanism

Fig. 12. Graphical Analysis of Current Video Tracking Implementation.

From Fig. 12, it can be seen that there is less survey work in this area, as well, as more emphasis is given to the conventional mechanism of point-based and kernel-based tracking system. Contribution towards silhouette-based is very few to find. Apart from this, the number of journal publications towards kernel-based is significantly less as compared to point-based tracking. This eventually shows that there was no equal emphasis being given to all the taxonomies of the video tracking system.

B. *Research Gap*

Different variants of research work are being carried out towards the video tracking system with a unique focus on accuracy. Every implementation offers a productive guideline towards adopting an effective methodology towards addressing the problems while also associated with a specific limitation and issues. Following are the list of open-end research issues which demands attention:

- Less Simplified Feature Extraction Process: Apart from extracting unique features, it is essential to ensure cost-effective modeling adherence. The majority of the existing approaches are highly inclined towards extraction of local level features, limiting the applicability in case of a change in visual and scene context. There is an emergent need to include a global level of features, which should result in inclusive of both low and high levels of attributes towards facilitating effective modeling of the video tracking system.

- Less Focus on Processing Time: An effective video tracking algorithm and system will definitely demand almost instantaneous response time. Without this inclusion, the practicability cannot be defined precisely. The existing system uses iterative and complete modeling towards a video tracking system due to its sole focus on achieving accuracy in its performance. With the inclusion of different challenges like different variants of occlusion, multi-view tracking, and sophistication of algorithm operation, the system must offer an instantaneous response in the presence of any dynamic video sequences.

- Need to Emphasize on Dimension Reduction: While performing video tracking, the system undergoes extraction of various informative contents, which are required to be stored and processed for improving preciseness in tracking. This is the case mainly with learning-based algorithms, which demands a higher dimension of trained data. The inclusion of a higher dimension of trained data will increase the memory complexity and increase the processing time to yield an appropriate response. Hence, there is a need to evolve up with an approach that can offer a better form of dimension reduction of the features considering the cases of a complex form of imageries in the video sequence, e.g. ariel images. There is also less focus on the optimization-based approach, which has good potential to deal with this open-end problem.

- The need to include contextual scene information: Existing approaches are built primarily over object detection followed by tracking. In the process of detection, the emphasis is only towards the foreground object and less towards the contextual information of the object and given scene. Without the inclusion of a contextual-based approach, the video tracking will have a limited scope of operation when exposed to uneven and dynamic mobility of an object whose heuristics are not present in the ruleset or ground truth or even in the trained image. Hence, contextual information demands enhanced scope.

VII. Conclusion

Irrespective of archival of the work carried out towards video tracking system, there is no evidence of any standardized model that acts as a benchmarked factor. Therefore, this article presents a typical insight into the identified three video tracking classes which is frequently found to be used: point-based tracking, kernel-based tracking, and silhouette-based tracking. It also briefs all standard methods that are witnessed to be implemented in these three standard video tracking algorithms. However, a closer look into the existing approach will only exhibit that they all can be further classified into three more classes of approach, i.e., contour-based tracking, tracking using native geometric model, and representation of a target object. All these associated models and their accuracy strongly depend upon how accurate the process of object detection and recognition is over a challenging scene of a video sequence. The study also concludes that each of the three standard categories discussed in this paper has both advantages and limiting factors, which should be improved upon to come up with a novel and effective video tracking scheme.

Therefore, our future work will emphasize addressing the open-end research issues discussed in the prior section. To do this, the future direction of work will emphasize more on modeling global features for the extraction process, along with an emphasis on precision. The future work will also be in the direction of inclusion of policy to balance the demands of higher accuracy and optimal processing time, lacking in existing approaches. Finally, an optimization-based approach could be implemented to address the issues connected with computational complexity.

References

[1] G. F. Shidik, E. Noersasongko, A. Nugraha, P. N. Andono, J. Jumanto and E. J. Kusuma, "A Systematic Review of Intelligence Video Surveillance: Trends, Techniques, Frameworks, and Datasets," in IEEE Access, vol. 7, pp. 170457-170473, 2019.

[2] Karam M. Abughalieh, Belal H. Sababha, and Nathir A. Rawashdeh. 2019. A video-based object detection and tracking system for weight sensitive UAVs. Multimedia Tools Appl. 78, 7 (April 2019), 9149–9167.

[3] A. Gautam and S. Singh, "Trends in Video Object Tracking in Surveillance: A Survey," 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2019, pp. 729-733.

[4] S. Salehian, P. Sebastian and A. B. Sayuti, "Framework for Pedestrian Detection, Tracking and Re-identification in Video Surveillance System," 2019 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Kuala Lumpur, Malaysia, 2019, pp. 192-197.

[5] K. Jin, X. Xie, F. Wang, X. Han and G. Shi, "Human Identification Recognition in Surveillance Videos," 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shanghai, China, 2019, pp. 162-167.

[6] O. Appiah, J. B. Hayfron-Acquah and M. Asante, "Real-Time Motion Detection and Surveillance using Approximation of Image Pre-processing Algorithms," 2019 IEEE AFRICON, Accra, Ghana, 2019, pp. 1-8.

[7] Y. Li, X. Zhang, H. Li, Q. Zhou, X. Cao and Z. Xiao, "Object detection and tracking under Complex environment using deep learning-based LPM," in IET Computer Vision, vol. 13, no. 2, pp. 157-164, 3 2019.

[8] M. Buric, M. Ivasic-Kos and M. Pobar, "Player Tracking in Sports Videos," 2019 IEEE International Conference on Cloud Computing Technology and Science (CloudCom), Sydney, Australia, 2019, pp. 334-340.

[9] Y. Ji, W. Li, K. Feng, B. Xing and F. Pan, "Automatic video mosaicking algorithm via dynamic key-frame," in Journal of Systems Engineering and Electronics, vol. 31, no. 2, pp. 272-278, April 2020.

[10] S. Bullinger, C. Bodensteiner, M. Arens and R. Stiefelhagen, "3D Object Trajectory Reconstruction using Instance-Aware Multibody Structure from Motion and Stereo Sequence Constraints," 2019 IEEE Intelligent Vehicles Symposium (IV), Paris, France, 2019, pp. 466-473.

[11] Wang, S., Lu, H., & Deng, Z. (2019). Fast object detection in compressed video. In Proceedings of the IEEE International Conference on Computer Vision (pp. 7104-7113).

[12] Li, C., Dobler, G., Feng, X. and Wang, Y., 2019. TrackNet: Simultaneous Object Detection and Tracking and Its Application in Traffic Video Analysis. arXiv preprint arXiv:1902.01466.

[13] L. Jiao et al., "A Survey of Deep Learning-Based Object Detection," in IEEE Access, vol. 7, pp. 128837-128868, 2019.

[14] S. A. Ahmed, D. P. Dogra, S. Kar and P. P. Roy, "Trajectory-Based Surveillance Analysis: A Survey," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 7, pp. 1985-1997, July 2019.

[15] S. S. Abdul Rajjak and A. K. Kureshi, "Recent Advances in Object Detection and Tracking for High Resolution Video: Overview and State-of-the-Art," 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), Pune, India, 2019, pp. 1-9.

[16] E. Trucco and K. Plakas, "Video Tracking: A Concise Survey," in IEEE Journal of Oceanic Engineering, vol. 31, no. 2, pp. 520-529, April 2006.

[17] Guo, Y., Yang, D. and Chen, Z., 2019. Object tracking on satellite videos: A correlation filter-based tracking method with trajectory correction by Kalman filter. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(9), pp.3538-3551.

[18] Gupta, M., Behera, L., Subramanian, V.K. and Jamshidi, M.M., 2014. A robust visual human detection approach with UKF-based motion tracking for a mobile robot. IEEE Systems Journal, 9(4), pp.1363-1375.

[19] del Rincón, J.M., Makris, D., Uruñuela, C.O. and Nebel, J.C., 2010. Tracking human position and lower body parts using Kalman and particle filters constrained by human biomechanics. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 41(1), pp.26-37.

[20] Wang, H. and Nguang, S.K., 2016. Multi-Target Video Tracking Based on Improved Data Association and Mixed Kalman/$ H_ {\infty} $ Filtering. IEEE Sensors Journal, 16(21), pp.7693-7704.

[21] Yang, J., Zhao, W., Han, Y., Ji, C., Jiang, B., Zheng, Z. and Song, H., 2019. Aircraft tracking based on fully conventional network and Kalman filter. IET Image Processing, 13(8), pp.1259-1265.

[22] Bhat, P.G., Subudhi, B.N., Veerakumar, T., Laxmi, V. and Gaur, M.S., 2019. Multi-Feature Fusion in Particle Filter Framework for Visual Tracking. IEEE Sensors Journal, 20(5), pp.2405-2415.

[23] Wang, J., Wang, Y. and Wang, H., 2016. Adaptive appearance modeling with point-to-set metric learning for visual tracking. IEEE Transactions on Circuits and Systems for Video Technology, 27(9), pp.1987-2000.

[24] Zhang, S., Lan, X., Qi, Y. and Yuen, P.C., 2016. Robust visual tracking via basis matching. IEEE Transactions on Circuits and Systems for Video Technology, 27(3), pp.421-430.

[25] Ahmadi, Kaveh, and Ezzatollah Salari. "A novel Multiple Hypothesis Testing (MHT) scheme for tracking of dim objects." 2015 IEEE International Conference on Electro/Information Technology (EIT). IEEE, 2015.

[26] Kutschbach, T., Bochinski, E., Eiselein, V. and Sikora, T., 2017, August. Sequential sensor fusion combining probability hypothesis density and kernelized correlation filters for multi-object tracking in video data. In 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 1-5). IEEE.

[27] Sheng, H., Chen, J., Zhang, Y., Ke, W., Xiong, Z. and Yu, J., 2018. Iterative multiple hypothesis tracking with tracklet-level association. IEEE Transactions on Circuits and Systems for Video Technology, 29(12), pp.3660-3672.

[28] Sheng, H., Zhang, Y., Wu, Y., Wang, S., Lyu, W., Ke, W. and Xiong, Z., 2020. Hypothesis testing based tracking with spatio-temporal joint interaction modeling. IEEE Transactions on Circuits and Systems for Video Technology, 30(9), pp.2971-2983.

[29] Yoo, H., Kim, K., Byeon, M., Jeon, Y. and Choi, J.Y., 2016. Online scheme for multiple camera multiple target tracking based on multiple hypothesis tracking. IEEE Transactions on Circuits and Systems for Video Technology, 27(3), pp.454-469.

[30] Baheti, B., Baid, U. and Talbar, S., 2016, June. An approach to automatic object tracking system by combination of SIFT and RANSAC with mean shift and KLT. In 2016 Conference on Advances in Signal Processing (CASP) (pp. 254-259). IEEE.

[31] Feng, F., Wu, X.J. and Xu, T., 2017, September. Object tracking with kernel correlation filters based on mean shift. In 2017 International Smart Cities Conference (ISC2) (pp. 1-7). IEEE.

[32] Liu, Y., Jing, X.Y., Nie, J., Gao, H., Liu, J. and Jiang, G.P., 2018. Context-aware three-dimensional mean-shift with occlusion handling for robust object tracking in RGB-D videos. IEEE Transactions on Multimedia, 21(3), pp.664-677.

[33] SGSEPRI, Z.P. and SJTU, C.Z., 2020, May. A New Method of Intelligent Maintenance of Smart Girds via Mean-Shift-Tracking Image Collaborative Displays. In 2020 IEEE 3rd International Conference on Electronics Technology (ICET) (pp. 348-351). IEEE.

[34] Shu, Z., Liu, G., Xie, Z. and Ren, Z., 2017, October. Real time target tracking based on nonlinear mean shift and particle filters. In 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) (pp. 1-5). IEEE.

[35] Tan, W.C. and Isa, N.A.M., 2017, May. Single sperm tracking using intersect cortical model-mean shift method. In 2017 International Conference on Signals and Systems (ICSigSys) (pp. 221-226). IEEE.

[36] Wang, H., Zhang, X., Yu, L. and Wang, X., 2018, August. Research on Mean Shift tracking algorithm based on significant features and template updates. In 2018 IEEE International Conference on Mechatronics and Automation (ICMA) (pp. 1199-1203). IEEE.

[37] Chen, Z., Li, B., Tian, L.F. and Chao, D., 2017, June. Automatic detection and tracking of ship based on mean shift in corrected video sequences. In 2017 2nd International Conference on Image, Vision and Computing (ICIVC) (pp. 449-453). IEEE.

[38] Dardas, N.H. and Georganas, N.D., 2011. Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. IEEE Transactions on Instrumentation and measurement, 60(11), pp.3592-3607.

[39] Feng, P., Wang, W., Dlay, S., Naqvi, S.M. and Chambers, J., 2016. Social force model-based MCMC-OCSVM particle PHD filter for multiple human tracking. IEEE Transactions on Multimedia, 19(4), pp.725-739.

[40] Yuan, Q., Thangali, A., Ablavsky, V. and Sclaroff, S., 2010. Learning a family of detectors via multiplicative kernels. IEEE transactions on pattern analysis and machine intelligence, 33(3), pp.514-530.

[41] Sun, J., Zhang, S. and Zhang, L., 2016. Object Tracking With Spatial Context Model. IEEE Signal Processing Letters, 23(5), pp.727-731.

[42] Sun, X., Zhang, L., Wang, Z., Chang, J., Yao, Y., Li, P. and Zimmermann, R., 2018. Scene categorization using deeply learned gaze shifting kernel. IEEE transactions on cybernetics, 49(6), pp.2156-2167.

[43] Yin, Y., Wang, X., Xu, D., Liu, F., Wang, Y. and Wu, W., 2016. Robust visual detection–learning–tracking framework for autonomous aerial refueling of UAVs. IEEE Transactions on Instrumentation and Measurement, 65(3), pp.510-521.

[44] Guo, J., Xu, T., Jiang, S. and Shen, Z., 2018, October. Generating reliable online adaptive templates for visual tracking. In 2018 25th IEEE International Conference on Image Processing (ICIP) (pp. 226-230). IEEE.

[45] Huang, X., Xu, J., Tai, Y.W. and Tang, C.K., 2020. Fast Video Object Segmentation With Temporal Aggregation Network and Dynamic Template Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8879-8889).

[46] Lin, S.D. and Chen, T.Y., 2018, October. Video Tracking Based on Template Matching and Particle Filter. In 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) (pp. 1-5). IEEE.

[47] Mutsam, N. and Pernkopf, F., 2020. Tracking of a Gunning Jet Using Particle Filtering in Infrared Image Sequences. IEEE Transactions on Instrumentation and Measurement.

[48] Pei, W.Y., Yang, C., Meng, L.Y., Hou, J.B., Tian, S. and Yin, X.C., 2018. Scene video text tracking with graph matching. IEEE Access, 6, pp.19419-19426.

[49] Rehman, B., Hong, O.W. and Hong, A.T.C., 2018, July. Using margin-based region of interest technique with multi-task convolutional neural network and template matching for robust face detection and tracking system. In 2018 2nd International Conference on Imaging, Signal Processing and Communication (ICISPC) (pp. 14-18). IEEE.

[50] Su, F., Fang, G. and Zou, J.J., 2018, July. Robust real-time object tracking using tiered templates. In 2018 13th World Congress on Intelligent Control and Automation (WCICA) (pp. 659-663). IEEE.

[51] Xiu, C. and Pan, X., 2017, May. Tracking algorithm based on the improved template matching. In 2017 29th Chinese Control And Decision Conference (CCDC) (pp. 483-486). IEEE.

[52] Kim, J., Choe, Y. and Kim, Y., 2011, September. High-quality 2D to 3D video conversion based on robust MRF-based object tracking and reliable graph-cut-based contour refinement. In ICTC 2011 (pp. 360-365). IEEE.

[53] Kishore, P.V.V., Prasad, M.V.D., Kumar, D.A. and Sastry, A.S.C.S., 2016, February. Optical flow hand tracking and active contour hand shape features for continuous sign language recognition with artificial neural networks. In 2016 IEEE 6th international conference on advanced computing (IACC) (pp. 346-351). IEEE.

[54] Luo, T., Chung, R.H. and Chow, K.P., 2014, March. A Novel object segmentation method for silhouette tracker in video surveillance application. In 2014 International Conference on Computational Science and Computational Intelligence (Vol. 1, pp. 103-107). IEEE.

[55] Pokheriya, M. and Pradhan, D., 2014, May. Object detection and tracking based on silhouette based trained shape model with Kalman filter. In International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014) (pp. 1-4). IEEE.

[56] Zou, T., Tang, X. and Song, B., 2011, November. Improved Camshift tracking algorithm based on silhouette moving detection. In 2011 Third International Conference on Multimedia Information Networking and Security (pp. 11-15). IEEE.

AUTHORS' PROFILE

Karanam Sunil Kumar Holds The Bachelor of Engineering in Computer Science and Engg. along with M. Tech Degree From VTU Belagavi.

Dr. NP Kavya holds Bachelor of Engineering in Computer Science and Engg. Along with MS in software systems and Ph. D in computer science from VTU Belagavi. She has vast experience of 24 years in the field of education and research. She is currently a Professor in Computer science and Engg. Department, RNSIT, Bengaluru. She has published around 90 research papers in reputed international journals including IEEE, Elsevier, Springier (SCI and Web of Science). Has 94+ citations in Google scholar as on Oct 2019. Her main areas of expertise are Machine Learning, Artificial Intelligence, and Big Data analytics.

Algorithm Design to Determine Possibility of Student Graduate Time in Student Grade Recapitulation Application

Marliana Budhiningtias Winanti[1], Umi Narimawati[2], Suwinarno Nadjamudin[3], Hendar Rubedo[4], Syahrul Mauluddin[5]*
Department Information System, Universitas Komputer Indonesia, Bandung, Indonesia[1]
Department Management, Universitas Komputer Indonesia, Bandung, Indonesia[2]
Department Informatic Engineering, International Woman University, Bandung, Indonesia[3]
Department Business Administration, International Woman University, Bandung, Indonesia[4]
Department Informatic Management, Universitas Komputer Indonesia, Bandung, Indonesia[5]

***Abstract*—This study aims to create an algorithm model to determine the potential time for student graduation to be applied to the grade recapitulation information system at XYZ University Information System Study Program. The XYZ University Information System Study Program already has a grade recapitulation information system, but the grade recapitulation information system has not been able to provide potential information about when a student can graduate from college. Information about probability graduating from college is very important as evaluation material in providing direction to students as an effort to achieve graduate on time. The more students who graduate on time, it can help increase the value of accreditation. The existing grade recapitulation information system can only display a history of grades and courses that have been taken by a student, so that the guardian lecturer has difficulty checking the courses that have not been taken by the student guidance and difficulty obtaining information when the student's guidance can graduate from college. In this study, an algorithm model was used to calculate the student's graduate time based on the calculation and mapping of subjects that had not been taken and had not yet passed. Based on the test results, the average time needed to determine student graduation time is 0.165 seconds.**

***Keywords*—Algorithm; graduate; subject; grade**

I. Introduction

Nowadays graduate on time is one of the most important concerns in higher education [1], so procedure or policy are needed to give a significant impact on graduation on time [2]. Graduating on time is one of the aspects of accreditation assessment in Indonesian tertiary institutions, especially in assessing the effectiveness and productivity of education. The more students graduate on time, the better the educational effectiveness and productivity of the study program will be.

The information systems study program at XYZ University in Indonesia has a relatively low graduate rate on time. The annual passing rate on time is below 50%. The low level of graduation time is due to the absence of supervision, especially for students who have the prospect to graduate overdue. The information system study program already has an executive information system that can provide information on the number of alumni who graduate on time and late graduate, but cannot provide information on active students who have the potential to graduate not on time. Information on the potential for graduating from college is very important as evaluation material in providing direction to students to graduate on time.

Based on the aforementioned problems, to increase the number of students who graduate on time, there is a need for a monitoring mechanism for students, especially students who have the probability to graduate not on time so that the guardian lecturers can provide direction and strategy in achieving graduate on time.

The purpose of this research is to design an algorithm to determine the possibility of students to graduate based on courses that have not been taken or have not passed. Research to determine the probability of graduate from college has been carried out by many other researchers through application of predictive algorithms such as the chaid method, naïve Bayes, decision tree, SVM algorithm, random forest, logistic regression and artificial neural network algorithms. However, in contrast to this research, this research will create an algorithm to calculate the possibility of graduate from college based on data subjects' grade that have not passed or have not been taken. Not based on a database of graduated student's grade. The research will provide information on how many more semesters must be taken, what courses can be taken each semester and whether they have the potential to graduate on time or not on time.

II. Literature Review

A. Related Work

Previous researchers have conducted several studies related to problem of graduate on time. Among them are those who conduct research to determine the factors inhibiting students from graduate on time and some are conducting research to predict graduation on time. However, based on the search and review, result of existing research has not been found similar to current research.

In 2013, Ni Komang Deby Ariani et al. Conducting research to determine the inhibit factors for graduates on time based on gender, major, entry route, origin area, origin school,

*Corresponding Author

and duration of final assignment. The conclusion of this study is that student graduation based on GPA with satisfactory category, it is found that there are no factors (independent variables) that affect student graduation time, while for student graduate based on GPA with very satisfying category, the factors that influence the time of graduation are gender, department, origin area, and duration of final assignment. Furthermore, for student graduates based on the GPA with the honors category, the factors that influence are origin area and the length of work on the final project [3]. In 2017, Widarto conducted research to determine the inhibiting factors seen from the problems in the final project process. Can be concluded that the cause of not graduate on time is the long distance to campus for guidance, still attending college, already have a job, less active lecturers, less conducive housing, and less intensive guidance [4].

To predict student graduation, many researchers apply data mining techniques with chaid method [5], naïve bayes [6], decision tree [7], SVM algorithm [8][9], random forest [10], artificial neural network algorithms [11], etc. Those studies using grades of graduate students then based on the data mining process the prediction results of student graduation are obtained. In contrast to this research, the prediction of graduates from college does not use data mining techniques that use data on alumni grades. However, in this study, the probability of graduation is calculated from the remaining subjects that have not graduated or have not been taken by an active student.

III. Methods

Stages of this research refer to the two stages of the prototype model system development method as shown in Fig. 1.

1) User requirement. This stage is the stage of collecting the data needed for the system/application by analyzing existing applications.

2) This stage is designing algorithms and conducting simulations by creating a feature to calculate the potential time of students' graduation in the student grade recapitulation application.

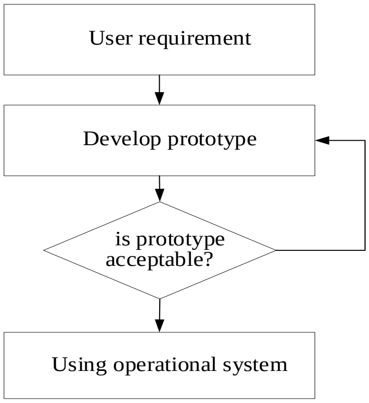


Fig. 1. Prototype Model [12].

IV. Results

At this stage, an analysis of the guardianship system procedure is carried out. The description of the guardianship process in the ongoing system is described in Fig. 2.

By knowing the guardianship procedure in the current system, it is known that guardian students and lecturers only use the history of course scores as material for course planning, thus allowing for errors in taking courses because they are not supported by a list of courses that have not been taken. In addition, the guardian lecturer cannot see the data of students who have the potential to graduate not on time so that in general there is no attention and effort to direct students to graduate on time.

The second stage is designing an algorithm to determine the potential time of student graduation and making an application for simulation or testing of the algorithm. In making the algorithm, considering the database structure of the running system. The current system database structure can be seen in Fig. 3.

The results of algorithm to calculate possibility of student graduation time are as follows:

1. Find student grade data in the database based on student ID. Store / display index data (code, subject, credit, semester, index, weight) in table / variable index while (T1)
2. Search curriculum data according to student data. Save / display curriculum data in tables / variables of temporary subjects (T2) (code, subject, credit, semester, index, weight)
3. Fill in the index and weight attributes in table T2 from table T1
4. Filter subjects with an E or Null index in table T2 and save it to the odd semester index table (T3) and even semester value table (T4)
5. Create four tables / variables (T5a-T5d) to store the distribution data for odd semester courses and create four tables / variables (T6a-T6d) to store distribution data for even semester courses. Allocate subjects in table T3 to odd semester tables (T5a-T5d) and table T4 to even semester tables (T6a-T6d) and adjust them to the number of credits.
6. Count the number of tables containing the subjects in T5a-T5d and T6a-T6d to find out the number of semesters that still have to be taken.
7. Display information about potential graduation time and information about whether or not to graduate on time.

The following is an example of applying the algorithm to calculate the potential time student to graduation.

1) Search student grade data in the database based on student ID. Save / Display value data (code, courses, credit, semester, Index, weight) in the temporary value table / variable (T1).

To accommodate the search result value data, prepare a 2 - dimensional array variable or table component (such as the table component in java SE) with a structure as in Table I.

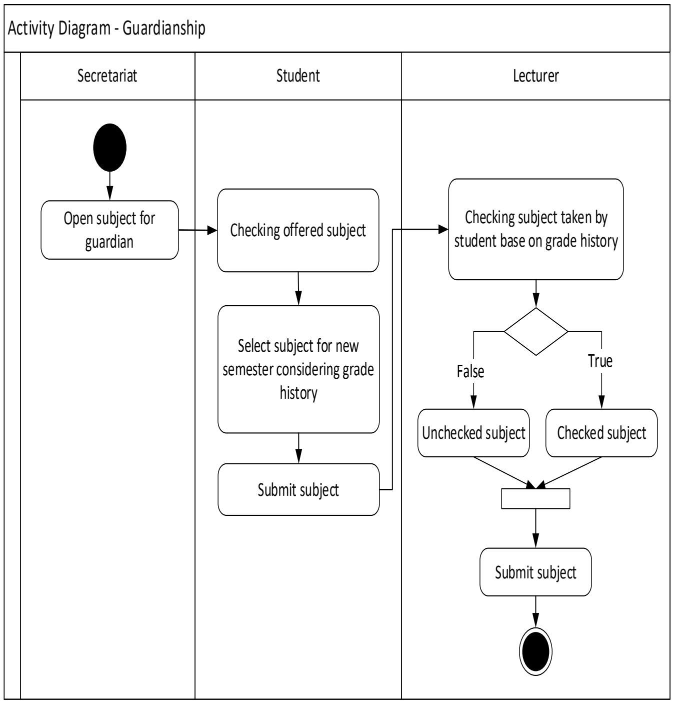

Activity Diagram - Guardianship
Secretariat
Student
Lecturer
Open subject for guardian
Checking offered subject
Select subject for new semester considering grade history
Submit subject
Checking subject taken by student base on grade history
False
True
Unchecked subject
Checked subject
Submit subject

Fig. 2. Guardianship Procedure.

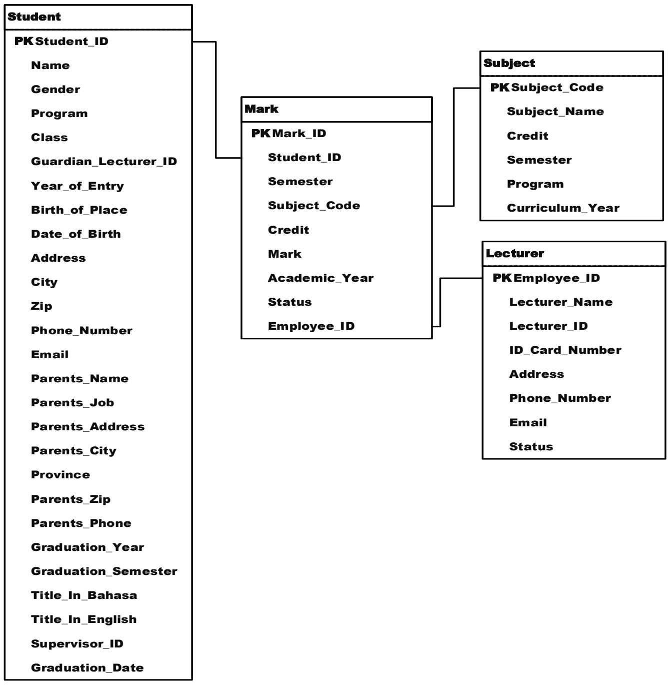


Fig. 3. Database of Grade Recapitulation Application [13].

2) Search curriculum data according to student data. Save / display curriculum data in tables / variables of temporary subjects (T2) (code, subject, credit, semester, index, weight)

After the first process is complete, continue by taking the curriculum data then store / display it with a structure like in stage 1, but the Index and weight are empty. Example of storing / displaying course data as in Table II.

TABLE I. EXAMPLE OF TABLE/VARIABLE TEMPORARY GRADE (T1)

No.	Code	Subject	Credit	Semester	Index	Weight
1	001	Subject 1	3	1	A	4
2	002	Subject 2	3	1	E	0
3	003	Subject 3	2	1	B	3
…						
7	007	Subject 7	3	2	E	0
8	008	Subject 8	3	2	C	2
9	009	Subject 9	3	2	D	1
…						

TABLE II. EXAMPLE OF TABLE/VARIABLE TEMPORARY SUBJECT (T2)

No.	Code	Subject	Credit	Semester	Index	Weight
1	001	Subject 1	3	1		0
2	002	Subject 2	3	1		0
3	003	Subject 3	2	1		0
..						
7	007	Subject 7	3	2		0
8	008	Subject 8	3	2		0
9	009	Subject 9	3	2		0
…						
14	014	Subject 14	2	3		0
15	015	Subject 15	3	3		0
16	016	Subject 16	3	3		0
…						

3) Fill in the index and weight attributes on T2 from T1: After the value data (T1) and subject data (T2) are available in the temporary variables / tables, do the process of filling in the index and weight in table T2 from the index and weight in table T1. Examples of filling in index and weight are as in Table III.

TABLE III. FILL IN INDEX AND WEIGHT AT T2

No.	Code	Subject	Credit	Semester	Index	Weight
1	001	Subject 1	3	1	A	4
2	002	Subject 2	3	1	E	0
3	003	Subject 3	2	1	B	3
..						
7	007	Subject 7	3	2	E	0
8	008	Subject 8	3	2	C	2
9	009	Subject 9	3	2	D	1
…						
14	014	Subject 14	2	3		0
15	015	Subject 15	3	3		0
16	016	Subject 16	3	3		0
…						

4) Filter subjects with an E or Null value in table T2 and save it to the odd semester value table (T3) and even semester value table (T4).

The next stage is to prepare two new tables / variables, namely the odd semester table (T3) and the even semester table (T4) to store / display subjects with E or Null index. Separate subjects worth E or Null according to the semester. Save odd semester subjects to table T3 and even semester subjects save to table T4. An example of the results of stage 4 is in Tables IV and V.

5) Make four tables / variables (T5a-T5d) to store the distribution data for odd semester courses and create four tables / variables (T6a-T6d) to store distribution data for even semester subjects. Allocate subjects in table T3 to odd semester tables (T5a-T5d) and table T4 to even semester tables (T6a-T6d) and adjust them to the number of credits.

At this stage prepare four tables / variables (T5a-T5d) to store the odd semester Subjects and four tables / variables (T6a-T6d) to store even semester subjects. In this study, the rules for allocating subjects are based on the order of the smallest to the largest semesters, the credit package quota is met. An example of the results of this stage can be seen in Tables VI to XIII.

6) Count the number of tables filled in odd and even semesters to determine the number of semesters that still have to be completed.

TABLE IV. EXAMPLE OF TABLE OF SUBJECT WITH E INDEX AND NULL IN ODD SEMESTER (T3)

No.	Code	Subject	Credit	Semester	Index	Weight
1	002	Subject 2	3	1	E	0
2	014	Subject 14	3	3		0
3	015	Subject 15	3	3		0
4	016	Subject 16	3	3		0
…						
13	051	Subject 51	2	7		0
14	052	Subject 52	3	7		0
15	053	Subject 53	3	7		0

TABLE V. EXAMPLE OF TABLE WITH SUBJECT WITH INDEX E AND NULL IN EVEN SEMESTER (T4)

No.	Code	Subject	Credit	Semester	Index	Weight
1	007	Subject 7	3	2	E	0
2	023	Subject 23	3	4		0
3	024	Subject 24	3	4		0
4	025	Subject 25	3	4		0
…						
18	057	Subject 57	3	8		0
19	058	Subject 58	2	8		0
20	059	Subject 59	3	8		0

TABLE VI. 1ST ODD SEMESTER TABLE (T5A)

No.	Code	Subject	Credit	Semester	Index	Weight
1	002	Subject 2	3	1	E	0
2	014	Subject 14	3	3		0
3	015	Subject 15	3	3		0
4	016	Subject 16	3	3		0
5	030	Subject 30	2	5		0
6	031	Subject 31	3	5		0

TABLE VII. 2ND ODD SEMESTER TABLE (T5B)

No.	Code	Subject	Credit	Semester	Index	Weight
1	033	Subject 33	3	5		0
2	034	Subject 34	3	5		0
3	035	Subject 35	3	5		0
4	047	Subject 47	2	7		0
5	048	Subject 48	3	7		0
6	049	Subject 49	3	7		0

TABLE VIII. 3RD ODD SEMESTER TABLE (T5C)

No.	Code	Subject	Credit	Semester	Index	Weight
1	051	Subject 51	2	7		0
2	052	Subject 52	3	7		0
3	053	Subject 53	3	7		0

TABLE IX. 4TH ODD SEMESTER TABLE (T5D)

No.	Code	Subject	Credit	Semester	Index	Weight
1						
2						
…						

TABLE X. 1ST EVEN SEMESTER TABLE (T6A)

No.	Code	Subject	Credit	Semester	Index	Weight
1	007	Subject 7	3	2	E	0
2	023	Subject 23	3	4		0
3	024	Subject 24	3	4		0
4	025	Subject 25	3	4		0
5	026	Subject 26	2	4		0
6	027	Subject 27	3	4		0
7	028	Subject 28	3	4		0

Based on the example of Tables VI to XIII, it can be counted that there are six semesters whose table contains the subject. So, it can be concluded that the number of semesters that must be taken is six semesters.

TABLE XI. 2ND EVEN SEMESTER TABLE (T6B)

No.	Code	Subject	Credit	Semester	Index	Weight
1	038	Subject 38	3	6		0
2	039	Subject 39	3	6		0
3	040	Subject 40	3	6		0
4	041	Subject 41	3	6		0
5	042	Subject 42	2	6		0
6	043	Subject 43	3	6		0
7	044	Subject 44	3	6		0

TABLE XII. 3RD EVEN SEMESTER TABLE (T6C)

No.	Code	Subject	Credit	Semester	Index	Weight
1	054	Subject 54	3	8		0
2	055	Subject 55	3	8		0
3	056	Subject 56	3	8		0
4	057	Subject 57	3	8		0
5	058	Subject 58	2	8		0
6	059	Subject 59	3	8		0

TABLE XIII. 4TH EVEN SEMESTER TABLE (T6D)

No.	Code	Subject	Credit	Semester	Index	Weight
1						
2						
…						

7) Display information about potential graduation time and information on whether to graduate on time or not on time.

To find out the potential time for graduating from college and information on whether to graduate on time or not on time, it can be done by adding the six semesters to the current semester position. If now in semester 3 then 3 + 6 is 9, meaning that the potential time to pass is in semester 9. Because the graduation time is in the 9th semester, the information displayed is not graduating on time because the rule on time is if you graduate in semester 8.

That's the algorithm design stage to calculate the probability of students to graduate from college. The next step is to create a feature to calculate the potential time for graduating from college in the student grade recapitulation application using the algorithm. Then the test is carried out to calculate the length of the function process to calculate the potential time to graduate students by inserting the function timer counter code.

The following are the results of testing on the algorithm applied to calculate the potential time of graduate students in 10 times, can be seen in Table XIV.

TABLE XIV. ALGORITHM TESTING RESULT

Test Time	Time (in Second)
1	0,187
2	0,155
3	0,160
4	0,139
5	0,182
6	0,182
7	0,137
8	0,158
9	0,202
10	0,150
average	0,165

V. CONCLUSION

Based on the simulation results of the algorithm implementation, information on the number of semesters that must be taken to graduate from college, the courses that can be taken in each semester and information on the possibility for passing on time or not on time. And based on testing by applying the prototype application to produce this information, it takes an average layout time in 0.165 seconds.

ACKNOWLEDGMENT

This research project is supported by Directorate of Research and Community Service, Ministry of Education and Culture of Indonesia in 2020 Fiscal Year.

REFERENCES

[1] H. Yue and X. Fu, “Rethinking Graduation and Time to Degree: A Fresh Perspective,” Res. High. Educ., vol. 58, no. 2, pp. 184–213, 2017.

[2] K. S. Lockeman and L. E. Pelco, “The Relationship between Service-Learning and Degree Completion,” Michigan J. Community Serv. Learn., vol. 20, no. 1, pp. 18–30, 2013.

[3] N. K. D. Ariani, I. W. Sumarjaya, and T. Bagus Oka, “Analisis Faktor-Faktor Yang Memengaruhi Waktu Kelulusan Mahasiswa Dengan Menggunakan Metode Gompit (Studi Kasus: Mahasiswa Fakultas MIPA Universitas Udayana),” E-Jurnal Mat., vol. 2, no. 3, p. 40, Aug. 2013.

[4] W. Widarto, “Faktor Penghambat Studi Mahasiswa yang Tidak Lulus Tepat Waktu di Jurusan Pendidikan Teknik Mesin FT UNY,” J. Din. VOKASIONAL Tek. MESIN, vol. 2, no. 2, p. 127, Oct. 2017.

[5] I. A. S. Padmini, L. P. Suciptawati, and M. Susilawati, “Analisis Waktu Kelulusan Mahasiswa Dengan Metode Chaid (Studi Kasus: Fmipa Universitas Udayana),” E-Jurnal Mat., vol. 1, no. 1, 2012.

[6] M. W. Amelia, A. S. M. Lumenta, and A. Jacobus, “Prediksi Masa Studi Mahasiswa dengan Menggunakan Algoritma Naïve Bayes,” J. Tek. Inform., vol. 11, no. 1, 2017.

[7] A. Rakhman, “Prediksi Ketepatan Kelulusan Mahasiswa Menggunakan Metode Decision Tree Berbasis Particle Swarm Optimation (PSO),” Smart Comp Jurnalnya Orang Pint. Komput., vol. 6, no. 1, pp. 193–197, 2017.

[8] M. I. Zulfa, A. Fadli, and Y. Ramadhani, “Classification model for graduation on time study using data mining techniques with SVM algorithm,” in AIP Conference Proceedings, 2019, vol. 020006, no. 2019.

[9] V. Riyanto, A. Hamid, and R. Ridwansyah, "Prediction of Student Graduation Time Using the Best Algorithm," Indones. J. Artif. Intell. Data Min., vol. 2, no. 1, pp. 1–9, Mar. 2019.

[10] A. S. Hoffait and M. Schyns, "Early detection of university students with potential difficulties," Decis. Support Syst., vol. 101, pp. 1–11, 2017.

[11] J. S. Bassi, E. G. Dada, A. A. Hamidu, and M. D. Elijah, "Students Graduation on Time Prediction Model Using Artificial Neural Network," IOSR J. Comput. Eng., vol. 21, no. 3, pp. 28–35, 2019.

[12] R. McLeod and G. Schell, Sistem Informasi Manajemen. Jakarta: Indeks, 2004.

[13] S. Mauluddin and R. Sidik, "Reverse Engineering in Student Mark Recapitulation Application," in IOP Conference Series: Materials Science and Engineering, 2019, vol. 662, no. 2.

Hybrid Invasive Weed Optimization with Tabu Search Algorithm for an Energy and Deadline Aware Scheduling in Cloud Computing

Pradeep Venuthurumilli[1]
Research Scholor, Acharya Nagarjuna University
Guntur, Andhra Pradesh

Dr. Sridhar Mandapati[2]
Dept.of Computer Applications, R.V.R and J.C College of Engg., Chowdavaram, Guntur-522019, Andhra Pradesh

***Abstract*—The current existing high flexibility, profitability, and potential have made cloud computing extremely popular among the companies. This is used for improving and applying resources in an efficient manner and optimize makespan of the tasks. Scheduling is easy while there are only a few tasks to complete with few resources. Contrastingly, at the time the users forward several demands to the environment of the cloud, there may be a need for optimally selecting and allocating resources for achieving the desired quality of service that makes scheduling challenging. In this work, using intelligent metaheuristic algorithms for processing the requests and tasks of users in energy-aware scheduling made for a deadline is proposed. Genetic Algorithm (GA) the evolutionary algorithm that is inspired by the natural process of selection and the evolution theory. The Invasive Weed Optimization (IWO) was yet another novel stochastic based on the population that was a derivative-free technique of optimization inspired by the growth of the weed plants. The TABU Search (TS) was a generalization technique of local search where the TABU list was used for preventing cycling and further generating the candidates of the neighborhood. A hybrid GA with the TS (GA-TS) with a hybrid IWO with TS (IWO-TS) has been proposed for the energy and deadline aware scheduling. The framework further offers optimization of energy and performance. The primary purpose of this algorithm has been to improve deadline and scheduling in cloud computing along with local as well as global search algorithms. This framework will offer optimization of performance and energy. The reason behind presenting this algorithm was improving both scheduling and deadline in cloud computing using both local and global algorithms and results proved the algorithm to have better results.**

Keywords—Cloud computing; scheduling; Genetic Algorithm (GA); Invasive Weed Optimization (IWO) and Tabu Search (TS)

I. Introduction

Cloud computing can be termed as a new paradigm where computing has been delivered as a new service as opposed to being a product with shared resources. A major advantage of cloud computing is the reduction of capital expenditure for cloud users and their service providers. It is well-suited for a range of applications like hosting websites, scientific workflow, high-performance computing, customer services and social networking [1].

Cloud computing may be termed as a model of expanding computation that is based on the technology of virtualization which is in response to the request of users through the network of the internet with dynamic resource allocation. Such virtualization will reduce the cost of maintenance of the organizations and also improve their accessibility. For the purpose of cloud computing, at the time a user asks for services, task scheduling becomes important for selection. The primary objective of such scheduling is to identify a new and optimal mapping for a task set and scheduling is quite easy at the time the number of resources and tasks is less. On opposing, at the time users send different demands to the environment of the cloud, there may be a need to choose optimally and allocate resources to the desired quality of users making scheduling challenging [2].

The problems in scheduling can be challenging as are combinatorial optimization problems which is either a maximization or a minimization issue containing: (1) one set of instances, (2) one finite set of the candidate solutions for every instance, and (3) a function for every instance. These problems will be solved only by finding an optimal solution to it. The problems are also Non-Deterministic Polynomial (NP)-hard and finding an efficient method for this is challenging. There are other problems in scheduling like timetable scheduling, processor scheduling, job-shop scheduling and so on [3].

Both “meta” and “heuristic” originated from Greek. The former refers to a “higher level” and the latter refers to “be aware” or “investigate”. To identify a good and optimal solution that has low cost of computation there are heuristic methods used that do not have any guarantee of feasibility or optimality. For enhancing efficiency, there is a set of metaheuristics that are employed for which there are numerous methods. Furthermore, the current algorithms will have lag in performance such as being trapped into a low optimum or a slow convergence rate. They may also have complex operators. Thus, presenting newer algorithms for catering to the weaknesses is a major and problematic aspect here [4].

Speaking in simple terms, metaheuristics may be a general class of an algorithm that is employed for finding solutions to problems of optimization at the time the normal techniques fail to help. By using the metaheuristic, an objective function can be of a type that precedes into considering different objectives. At the same time, all the other metaheuristic algorithms will have a suspicious modification of the parameters of optimization that can be crucial to discover better solutions that lack time. There are several metaheuristics that are available with some novel variations recommended for the allocation of resources in different fields. Some more metaheuristics are very prominent in cloud computing and for the management of resources. These may be the Memetic Algorithm (MA), League Championship Algorithm (LSA), Ant Colony Optimization (ACO), Immune Algorithm (IA), Cuckoo Search (CS), Harmony Search, (HS), Firefly Algorithm (FA) and the Differential Evolution (DE).

This hybrid optimization approach was proposed for designing a linkage method for its dimensional synthesis to merge the advantages of the stochastic and the deterministic methods of optimization. For this work, the GA, TS, IWO, hybrid GA-TS and the IWO-TS algorithm were proposed for the deadline scheduling that is energy-aware. The rest of this examination has been composed therefore. The related work in the writing has been examined in Section 2. All strategies utilized are clarified in Section 3. Results were presented in Section 4 and the end was made in Section 5.

II. Literature Works

For the purpose of this work, the problem of task scheduling was taken to be a problem that was bi-objective that includes the minimization of consumption of energy and its makespan. Firstly, Sahoo et al., [5] had proposed another novel Learning Automata-based Scheduling (LAS) based framework for the tasks that were deadline sensitive. Later, there was a LAS algorithm that had been introduced to exploit the task heterogeneity. There was some extensive simulation that was conducted for designating the applicability and effectiveness of the LAS for the task scheduling that was deadline sensitive in a cloud environment.

For the purpose of this work, there was a new cost-aware job scheduling method that was based on the queuing theory presented by Li et al., [6]. The problem of job scheduling in a private cloud was modelled to be a queuing model. After this, the time for task execution was predicted using a backpropagation neural network. A max–min strategy was applied to the schedule tasks regarding the outcome of forecast in the half and half mists. Its goal is utilizing assets appropriately and overseeing load between assets with least execution time. High correspondence cost brings about in mists forestall task schedulers from being applied in an enormous scope conveyed condition. Booking is a noticeable movement that is executed in a distributed computing condition. To build distributed computing remaining task at hand productivity, assignments planning is performed to get greatest benefit. In cloud, high correspondence cost forestalls task schedulers from being applied in huge scope dispersed conditions. Cloud condition framework planning is NP-finished. To explain the NP complete and NP difficult issues heuristic methodologies are utilized.

The experiment showed this cost-aware job scheduling algorithm to decrease the average response waiting time for a job in a private cloud. Also, the proposed algorithm of job scheduling was able to progress the throughput of the system with an average task response time and also the total cost in the hybrid clouds.

The metaheuristic techniques will have their efficiency established in a range of algorithms of workflow scheduling but it is yet not known if the metaheuristic selected is well-suited to solve it. Shishido et al., [7] had examined the effects of the GA and the PSO in the optimization of workflow scheduling. There was an algorithm of security and cost-aware workflow scheduling chosen for the evaluation of metaheuristic performance. There were three algorithms evaluated using real-world workflows having a risk limitation ranging between 0 and 1 and a 0.1 step. The GA-based algorithms had suggestively overtook the PSO in terms of response time and efficacy of cost.

Liu et al., [8] had recently modelled task scheduling in the form of an optimization problem that considered time deadline constrained and unconstrained cases. For addressing this issue, there was a Heterogeneous Earliest Finish Time (HEFT) with a technique for a request for inclination to perfect arrangements named as the HEFT-T calculation. For an unconstrained case, there was another three-phase approach which depended on the HEFT-T calculation to pick ideal arrangements utilizing the non-overwhelmed arranging approach. On account of a cutoff time obliged choice, there was a versatile weight alteration put together methodology with respect to the HEFT-T that was proposed for adjusting weight values in the case of time. In comparison with the other algorithms that were state-of-the-art, the proposed one has performed better in terms of optimization criterion for cost and mean load.

Since there are certain inherent defects found in mobile devices like limited space for storage of battery backup, there are some challenges faced in the management of mobility, security, energy management, and QoS. This has resulted in the emergence of several paradigms like fog computing and Mobile Cloud Computing (MCC). All these paradigms permit offloading certain tasks for the execution of the cloud that makes tasks scheduling important in the mobile cloud and the mobile device. Tang et al., [9] [19] have modelled the problem to be of optimization of energy consumption and this also considers the dependency of the task, transmission of data and other constraints like cost and time deadline.

Haidri et al., [10] [18] had addressed these problems found in the scheduling of the workflow tasks on the cloud computing systems like Total Price for Execution (TPE) and the Total Execution Time (TET). This is for the workflow and for meeting the limitations of a deadline in an environment that is stochastic. As opposed to the TET and the TPE, the acquisition delay of the which is a major trait of the cloud. The work initially formulates an issue of a model of stochastic scheduling on the cloud. There is a Stochastic Cost-Effective Deadline-Aware (S-CEDA) based resource scheduler that had been established. There was the S-CEDA that had incorporated the variance and the expected value of the processing time of the task and inter-task communication workflow scheduling.

III. Methodology

Distinguishing a worldwide minimizer is a difficult errand. The creator has misused both the highlights of GA, IWO and TS be the altered variant of a coordinated calculation. Capacity of worldwide hunt and getaway from a nearby least is improved and a mixture GA-TS and IWO-TS methods are examined.

A. Genetic Algorithm (GA)

The GA had been settled centered on the principle of "survival of the fittest" by Charles Darwin. It exploits the best solution that has been accomplished by the searches conducted earlier along with an investigation of newer sections of their explanation planetary. A solution for a given problematic has been signified by a new gene that contained elements known as genes. The algorithm will begin with a population of solutions made in a random fashion and the population's quality will be identified by means of evaluating them with any objective function called the fitness function. Once the population is initialized using solutions generated randomly, there are new descendants genes that are generated using either the crossover or the mutation operations. The solutions generated are then evaluated and acclimated for being adapted to the people if it is better. This process that generates offspring will be repeated until an optimal solution with the best offspring has been created [11][17].

B. Invasive Weed Optimization (IWO) Algorithm

The IWO had been initially brought about by Mehrabian and Locus which is an algorithm based on population originated from the weed colonial behaviour. The IWO is quite simple but very efficient in identifying optimum solutions to the objective function. This has been implemented on the basis of the features of the weeks found in a colony as reproduction, the competition to survive and growth. When compared to the other algorithms, the IWO is very simple and has sufficient convergence ability and rate. There are some more major features found in the algorithm that is different from the other methods and these are exclusive competition, distribution of space and reproduction. For simulating the weed behaviour, the following algorithm is used as per [12].

Stage 1: The Initial populace creation: each populace will require N0 seeds that are disseminated haphazardly inside an n-dimensional space.

Stage 2: Reproduction: each seed will develop and turn into a develop plant after which the seed creation happens for another new age. The seeds that are created by each plant will increment directly between two distinct estimations of the base (Smin) and the greatest (Smax) that have potential measures of seeds delivered. The genuine measure of seeds that are created for the ith plant for each and every recurrent will be founded on the objective worth (Fi), it's the best (Fbest) and the most exceedingly awful (Fworst) objective qualities registered according to condition (1):

$$Numseed(i) = \left[S_{\min} + (S_{\max} - S_{\min}) \frac{f - f_{worst}}{f_{best} - f_{worst}} \right] \tag{1}$$

Step 3: The distribution will space both assimilation and randomness of this algorithm that is connected to this particular stage. The seeds that are produced will be distributed within the d dimensional space using a normal distribution that has a zero mean with a different variance which is ($N(0,\delta_t)^d$). The seeds will have to be closer to the breeder plant. Even though there is a standard deviation which will decrease from its initial amount ($\sigma_{initial}$) to the final amount (σ_{final}) for every repeat, in all simulations, where a non-linear variation of the standard deviation as in (2):

$$\sigma_t = \left(\frac{T-t}{T} \right)^n \times (\sigma_{initial} - \sigma_{final}) + \sigma_{final} \tag{2}$$

As per equations 1 & 2, T will represent the maximum repeats that are related to (t) and this will be a factor of non-linear modulation. For this status, the seed position (S_j) for that of the ith plant (W_i) will be calculated as per (3):

$$S_j = W_i + N(0,\delta_t)^d, 1 <= j \le numseed(i) \tag{3}$$

Step 4: Exclusive competition: the actual number of plants that are formed by a quick reproduction can reach its maximum value which is ($W_{\max}$). Each plant will be allowed to produce seeds according to the method of reproduction. These seeds will be authorized to spreading within the search space along with a correspondence to the distribution space method at the time the seeds reach their position; after this, they form a new colony alongside the parent plants. Those members having less propriety will be deleted for the other members to be able to reach the maximum allowable value. All parent plants will associate alongside their children and the ones having the highest propriety in the group will be preserved and replacement will be permitted. This method of the crowd control mechanism is forced on the subsequent generations till such time it reaches its final period.

Step 5: If criterion is met, end if not repeat from Step 2.

C. Tabu Search (TS) Algorithm

The TS is yet another metaheuristic algorithm that is used to solve problems in combinatorial optimization like the problem of bin packing. For the purpose of escaping from their potential local optimum, the TS will make use of a local or a neighbourhood search for moving from solution x to solution y in an iterative manner. The termination criteria used commonly for this is to make use of a fixed and maximum number of iterations, the maximum number of iterations that are done without refining the objective function, the maximum limit on the time taken for searching and so on. When the local search is executed, certain regions may not get traversed and the TS will resolve this by modifying their neighbourhood structure for the solution. These solutions will be acknowledged to $N_t(x)$, which is the new neighbourhood and will be determined by using memory structures [13]. This search will progress by an iterative move from solution x to solution x' as in $N_t(x)$.

The TS will make use of random selection by default and for speeding up the process of search it also uses a “TABU List (Tabus)” that can restrict search space and avoid cyclic behaviour. This TABU list is of short-term memory consisting of solutions recently searched. The TS eliminates solutions found in the TABU list from $N_t(x)$. At times, the TABU list can also avoid solutions from taking certain types of moves (for example an item which was added to the bin that cannot be removed within the subsequent n moves). The chosen attributes in the recently visited solutions will be considered as “TABU-active”. The solutions consisting of TABU-active elements will be called the “TABU”. Such types of short-term memory are known as a “Recency-based” memory. The TABU list will contain attributes that are more effective in certain types of domains even though they may result in the creation of a new obstacle. If a particular attribute is noticeable to be a TABU, it may result in more than one of such solutions being a TABU. Certain solutions will have to be avoided and they may also be of great quality but not visited yet. For the purpose of mitigating this issue, “aspiration criteria” had been introduced: these will override the TABU state of a solution thus with the excluded solution in the set. The aspirations criterion used most often is accepting solutions that are better than the current ones.

D. Proposed Hybrid Genetic Algorithm (GA) – Tabu Search (TS) Algorithm (GA-Tabu)

Premature convergence is probably the primary issue in the GA and the weakness of the TS was its dependent on initial solutions and single-point modes of search. The GA is able to make available initial solutions of high quality for the TS and also has a fast speed of the search that is able to reward for the problem of speed of the TS. Furthermore, the TABU list is flexible and the TS aspiration criterion is able to help the GA in escaping from the local optima. For the proposed GA-TS method, the TS has been integrated with the GA as follows: a mutation operator based on the TS will replace its original operator as soon as the warning for prematurity is triggered. For judging if the process of search was trapped inside premature convergence, the index of prematurity was defined by computing the degree of similarity between both individuals as in (4):

$$I = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{(N-2)!*2!}{N!} * \frac{W}{L} \tag{4}$$

Wherein, I means the rashness list, N signifies the genuine include of people in the populace, $N!/((N-2)!*2!)$ speaks to the check of blends of the pairwise people in the populace, Wrepresents the real include of same qualities in similar areas for each two people in the whole populace and Lsignifies the real length of chromosomes.

The means in the half breed GA and TS strategy proposed are [14]:

Stage 1: Fix the boundaries like populace size (popsize), the greatest emphasis of a half breed GA and TS (iter_GATS), the most extreme cycle of the TS (iter_TS), the hybrid likelihood (pc), the transformation likelihood (pm), length of the TABU rundown (lt).

Stage 2: Create the popsize populace. Unravel each person for getting wellness esteem and relate them to achieve an underlying arrangement x.

Stage 3: Set n=1.

Stage 4: Create the new populace for resulting ages. The hereditary advancement utilizing three administrators that incorporates choice, hybrid, lastly change will be applied for the formation of posterity for the resulting populace. On the off chance that chi, I = 1, 2… the popsize will be for the new populace. A roulette wheel methodology will be utilized for playing out a determination.

Stage 5: Presently every person in the new populace chk, k=1, 2… with a popsize as the TS beginning arrangement is utilized. Rehash stages 6 to 8 for improving individual quality until the emphasis approaches popsize. Presently continue to stage 9.

Stage 6: Set j=1.

Stage 7: Distinguish all areas of the TS and its present arrangement. Pick the best non-TABU neighborhood fulfilling the targetmeasure.

Stage 8: Presently update the TABU rundown to check end condition and if this (j = iter_TS) is met of if all the basic squares have not exactly an activity, stop the TS and the ideal arrangement. Individual chk will be subbed by the ideal arrangement. If not, j = j + 1 and now return to stage 7.

Stage 9: If the end condition (n = iter_GATS) is met, stop the half and half GA and TS technique and yield ideal arrangement. If not, n = n + 1 and now move back to stage 4.

The flowchart for proposed hybrid GA and TS method [15] as shown in Fig. 1.

E. Proposed Hybrid Invasive Weed Optimization (IWO)-Tabu Search (TS) Algorithm

There have been several attempts made for improvement strategies of the IWO and the TS for practical applications. There was a hybrid strategy that was in combination with the TS and the IWO which has been introduced for the generation of test data that satisfies the criterion of decision coverage. In the hybrid IWO-TS algorithm proposed, the IWO was applied for searching the entire solution space for ensuring diversity of population and also for the TS that is applied to a local search for avoiding convergence to the local optimum. For every iteration of the IWO-TS, it will not carry out any competitive exclusion operator once a spatial dispersion operation is conducted but will check if every weed has been enhanced by the seeds. The week is not enhanced inside the g_1 (g_1> 0) and successive iterations, it may remain stagnated and placed into the taboo list. The plants in the neighborhood of the tabooed weeds will have lower levels of fitness that has to be removed from the present population [16].

Let W denote the current weed set, S_i the new set of seeds that are produced by the weed $W_i \in W$ in the current iteration, and TL denotes the taboo list. Now the current population of the plant may be shown as (5):

$$P = W \cup S_1 \cup S_2 \cup ..... \cup S_{|W|} \tag{5}$$

The plants to be removed using a tabooed solution T$_i$∈TL are shown as (6)

$$E_i = \{P_j \in P \mid \Delta(P_j, T_i) \le \delta_i; F(P_j) < F(T_i)\} \tag{6}$$

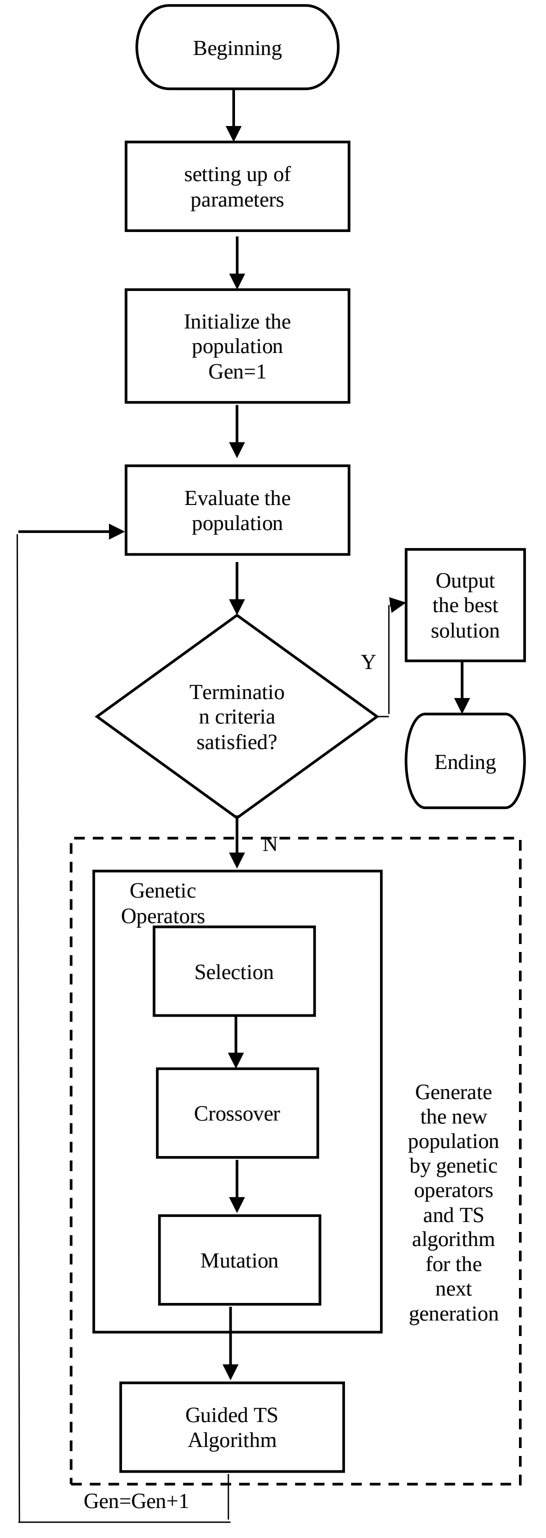


Fig. 1. Flowchart for the Proposed Hybrid GA-TS Algorithm.

Wherein, Δ (·,·) indicates the actual distance between two solutions, δ$_i$ denotes a positive parameter to decide the size of its tabooed regions using solution T$_i$, and F (·) which denotes its fitness function.

As both IWO-TS focus on the continuous problems, Δ (·,·) may be definite to be the Euclidean distance. Parameter δ$_i$ will have a significant level of influence on the IWO-TS and its performance. The higher the δ$_i$ is, the more will be the plants that are deleted by the T$_i$. It can be challenging to choose a fixed value for the different δ$_i$ for adapting to different problems. One more feasible way was to choose a value for every δ$_i$ to the solution and its real size. This is discovered by the IWO-TS around its corresponding weed. This means, if the weed W$_i$∈W has been judged as tabooed (without any loss of generality it is assumed to correspond to T$_i$∈ TL), I δ will be set to (7):

$$\delta_i = \max\{\Delta(S_{ij}, W_i) \mid \forall S_{ij} \in S_i\} \tag{7}$$

When an elimination process of the plant P is complete, all remaining plants $P^{'} \in P \mid (E_1 \cup E_2 \cup .... \cup E_{|TL|})$ will be sent to the operator of competitive exclusion. A TS approach will provide well-organized methods to handle this but it may not always be easy for the tabooed regions to have promising solutions. So, an aspiration criterion is introduced and the TL is set with a fixed length tl (tl> 0) to release the tabooed solutions. Once this is done, the neighborhood will further explore certain better solutions. In case no tabooed solution gets released, the IWO-TS will have to identify a better solution with the current weeds. Also, introducing the aspiration criterion can avoid an infinite increase to the TL thereby saving cost. As the IWO-TS works on local search, some seeds may be near the tabooed weeds and it may be redundant to do aremoval operation using a tabooed weed for all iteration. An easy way was to occasionally perform in each g$_2$ (g$_2$> 0) iterations to reduce computation cost.

For the purpose of enabling the exploration of a new solution by the IWO-TS a new self-production operator has to be defined. This will generate new weeds randomly and provide chances to explore undetected regions. The weeds generated can get excluded in the subsequent iteration by means of a competitive exclusion operator that can bring down its performance. So, the weeds randomly generated will be moved to the best weed found until now and the primary motive will be to move them to a region with promising solutions and is described mathematically in (8 and 9):

$$W_{new}^{0} = X_{\min} + rand(0,1).(X_{\max} - X_{\min}) \tag{8}$$

$$W_{new} = W_{new}^{0} + rand(0,1).(W_{gb} - W_{new}^{0}) \tag{9}$$

Wherein the X$_{max}$ and X$_{min}$ denote the upper bound and the lower bound on different variables that are optimized, W$_{gb}$ denotes the best weed that is found so far, and the rand (0, 1) will return a random number that is uniformly distributed within the interval (0, 1). Since the IWO-TS have a strong ability of exploration in an early stage of the search, it will have to complete a fine local search in a later stage and it is important to keep changing the weeds that are created by this self-production operator in other iterations. The number has been altered as in (10):

$$n_{sp} = \left\lfloor (1-4(t/t_{\max}-0.5)^2)n_{\max} \times 20\% \right\rfloor \tag{10}$$

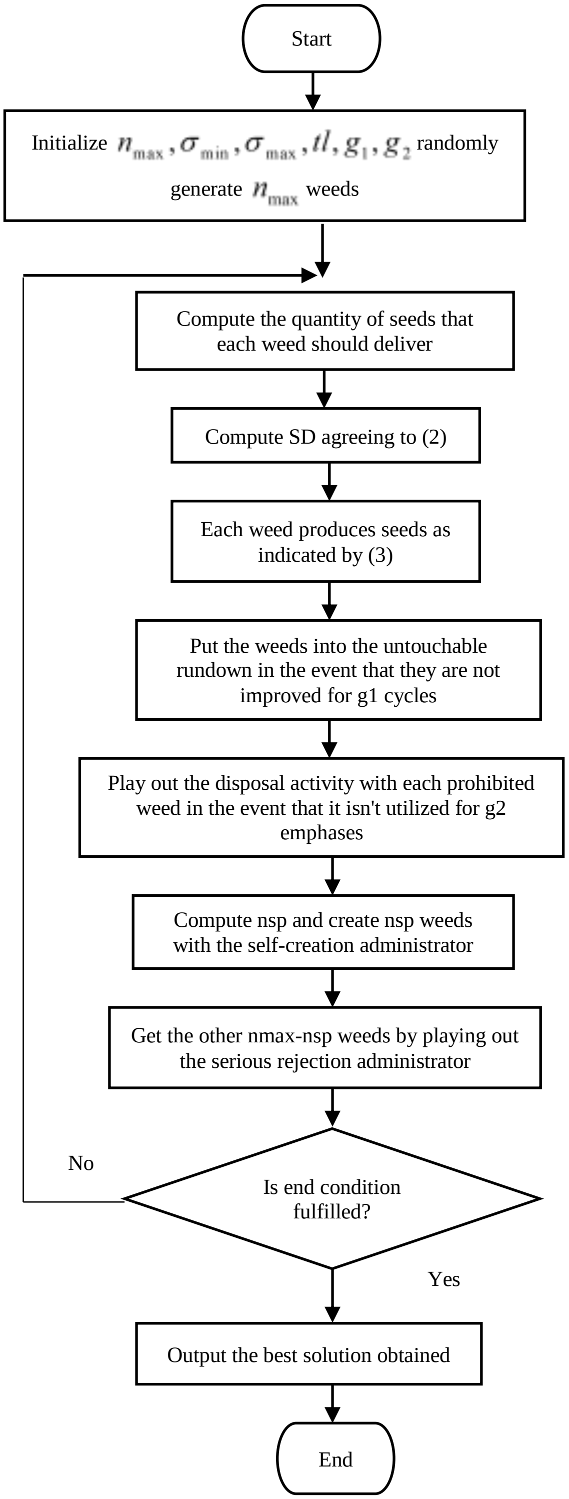

Start
Initialize $n_{\max}, \sigma_{\min}, \sigma_{\max}, tl, g_1, g_2$ randomly generate $n_{\max}$ weeds
Compute the quantity of seeds that each weed should deliver
Compute SD agreeing to (2)
Each weed produces seeds as indicated by (3)
Put the weeds into the untouchable rundown in the event that they are not improved for g1 cycles
Play out the disposal activity with each prohibited weed in the event that it isn't utilized for g2 emphases
Compute nsp and create nsp weeds with the self-creation administrator
Get the other nmax-nsp weeds by playing out the serious rejection administrator
Is end condition fulfilled?
No
Yes
Output the best solution obtained
End

Fig. 2. Flowchart for the Proposed Hybrid IWO-TS Algorithm.

Wherein, the notation $\lfloor \cdot \rfloor$ will be a round down operation. In accordance with (10), the n_{sp} will increase nonlinearly from zero till the search reaches the maximum value $\lfloor 0.2n_{\max} \rfloor$. After this, the n_{sp} will reduce gradually to zero. For every iteration, the IWO-TS will get another $n_{\max} - n_{sp}$ weeds from its plant set P' which is by carrying out on it a competitive exclusion operator.

The flowchart for hybrid IWO-TS as illustrated in Fig. 2.

IV. Results and Discussion

The GA, TS and IWO methods are used. Here, 5000 to 20000 number of tasks are considered. The guarantee ratio, energy savings and resource utilization are represented in Tables I to III and Fig. 3 to 5.

From the Fig. 3, it is seen that the IWO-Tabu has higher guarantee ratio by 3.24%, by same value, by 1.02%, by 2.06%, by same value, by same value and by same value for number of tasks 5000, 7500, 10000, 12500, 15000, 17500 and 20000 than GA-Tabu, respectively.

TABLE I. Guarantee Ratio % for IWO-Tabu

Number of Tasks	GA-Tabu	IWO-Tabu
5000	91	94
7500	98	98
10000	97	98
12500	96	98
15000	99	99
17500	99	99
20000	99	99

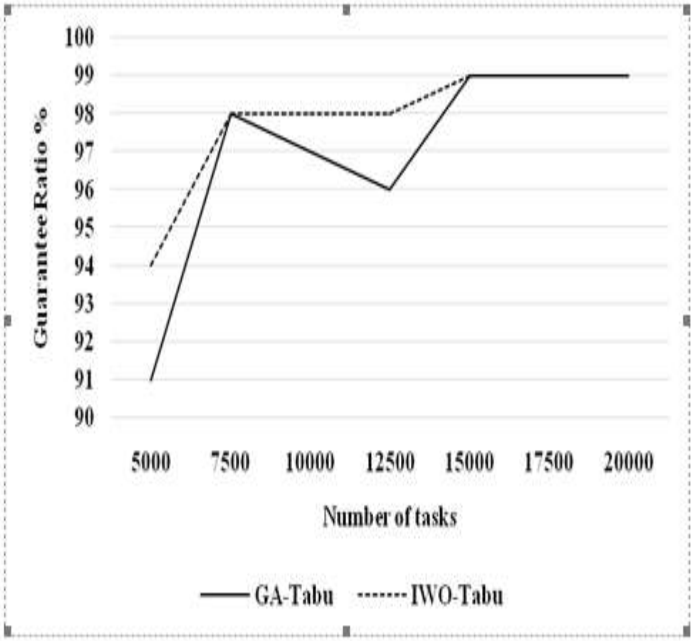


Fig. 3. Guarantee Ratio % for IWO-Tabu.

TABLE II. ENERGY SAVINGS % FOR IWO-TABU

Number of Tasks	GA-Tabu	IWO-Tabu
5000	3	3.3
7500	3.6	4.1
10000	3.3	4.4
12500	3.2	4.2
15000	3.6	4
17500	3.1	3.8
20000	2.8	3.3

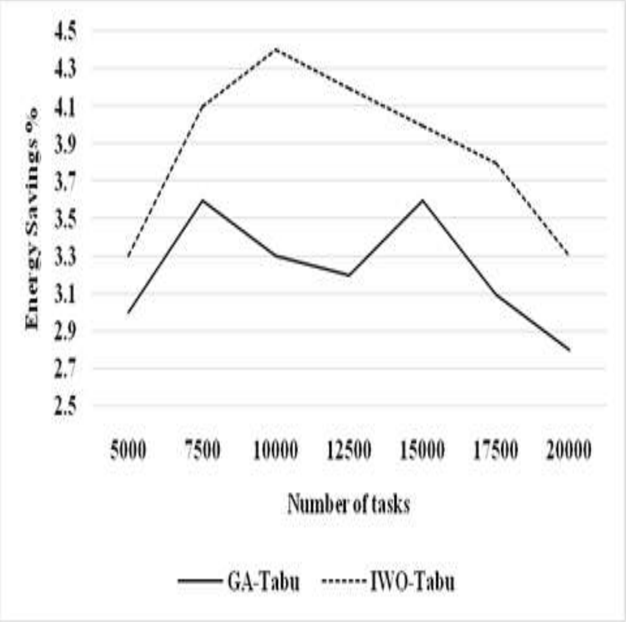


Fig. 4. Energy Savings % for IWO-Tabu.

From the Fig. 4, it is seen that the IWO-Tabu has higher energy savings by 9.52%, by 12.98%, by 28.57%, by 27.02%, by 10.52%, by 20.29%and by 16.39% for number of tasks 5000, 7500, 10000, 12500, 15000, 17500 and 20000 than GA-Tabu, respectively.

From the Fig. 5, it is seen that the IWO-Tabu has higher resource utilization by 4.58%, by 5.4%, by 3.59%, by 4.82%, by 3.51%, by 4.65%and by 6.97% for number of tasks 5000, 7500, 10000, 12500, 15000, 17500 and 20000 than GA-Tabu, respectively.

TABLE III. RESOURCE UTILIZATION % FOR IWO-TABU

Number of Tasks	**GA-Tabu**	**IWO-Tabu**
5000	64	67
7500	72	76
10000	82	85
12500	81	85
15000	84	87
17500	84	88
20000	83	89

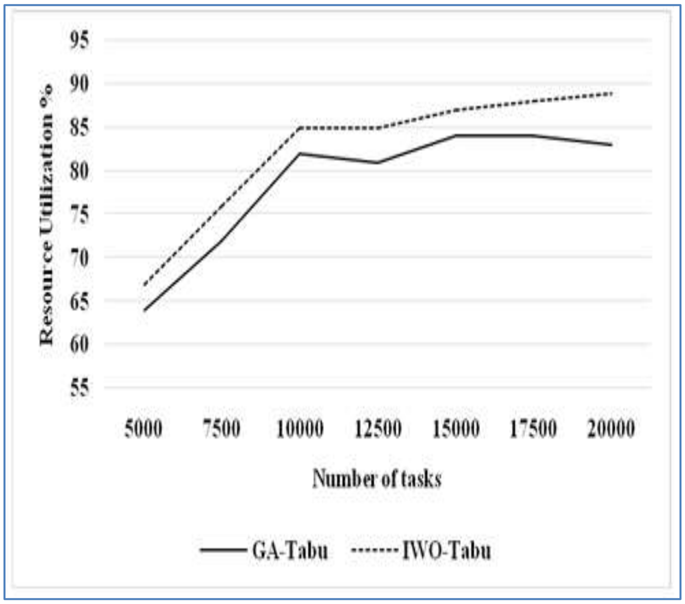


Fig. 5. Resource Utilization % for IWO-Tabu.

V. CONCLUSION

There is an ability to allocate limited resources for computing large tasks that have a goal of optimization inspiring wider solutions in the domain of cloud computing. To solve both algorithms like the GA or the TS (GA-TS) is used. Owing to the GA having the ability that handles complex search spaces that have a high chance of success to find optimal solutions. IWO is a metaheuristic that was recently developed there is a metaheuristic imitating the invasive weed behaviour. The reproduction and the spatial dispersal operators in the original IWO that make the seeds positioned in and around the best weed that can lead to premature convergence. For overcoming such drawbacks, the work proposed the IWO-TS algorithm using the core idea of the TS. If there is no better solution found, the IWO-TS will judge the weed that is stagnated and will taboo it avoiding any repeated search in the neighbourhood. Additionally, the IWO-TS will define another self-production operator to generate new weeds randomly as opposed to directly selecting it from the plant population in order to ensure new solutions are also explored.

REFERENCES

[1] Panda, S. K., & Jana, P. K. (2015). Efficient task scheduling algorithms for heterogeneous multi-cloud environment. The Journal of Supercomputing, 71(4), 1505-1533.

[2] Kashikolaei, S. M. G., Hosseinabadi, A. A. R., Saemi, B., Shareh, M. B., Sangaiah, A. K., &Bian, G. B. (2019). An enhancement of task scheduling in cloud computing based on imperialist competitive algorithm and firefly algorithm. The Journal of Supercomputing, 1-28.

[3] Sonawane, M. P. A., &Ragha, L. (2014). Hybrid genetic algorithm and TABU search algorithm to solve class time table scheduling problem. International Journal of Research Studies in Computer Science and Engineering, 1(4), 19-26.

[4] Madni, S. H. H., Latiff, M. S. A., Coulibaly, Y., &Abdulhamid, S. I. M. (2016). An appraisal of meta-heuristic resource allocation techniques for IaaS cloud. Indian Journal of Science and Technology, 9(4), 1-14.

[5] Sahoo, S., Sahoo, B., &Turuk, A. K. (2019). A Learning Automata-based Scheduling for Deadline Sensitive Task in The Cloud. IEEE Transactions on Services Computing.

[6] Li, C., Tang, J., & Luo, Y. (2018). Towards operational cost minimization for cloud bursting with deadline constraints in hybrid clouds. Cluster Computing, 21(4), 2013-2029.

[7] Shishido, H. Y., Estrella, J. C., Toledo, C. F. M., &Arantes, M. S. (2018). Genetic-based algorithms applied to a workflow scheduling algorithm with security and deadline constraints in clouds. Computers & Electrical Engineering, 69, 378-394.

[8] Liu, L., Fan, Q., & Buyya, R. (2018). A deadline-constrained multi-objective task scheduling algorithm in mobile cloud environments. IEEE Access, 6, 52982-52996.

[9] Tang, C., Xiao, S., Wei, X., Hao, M., & Chen, W. (2018, January). Energy Efficient and Deadline Satisfied Task Scheduling in Mobile Cloud Computing. In 2018 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 198-205). IEEE.

[10] Haidri, R. A., Katti, C. P., & Saxena, P. C. (2019). Cost-effective deadline-aware stochastic scheduling strategy for workflow applications on virtual machines in cloud computing. Concurrency and Computation: Practice and Experience, 31(7), e5006.

[11] Singh, P., Dutta, M., & Aggarwal, N. (2017). A review of task scheduling based on meta-heuristics approach in cloud computing. Knowledge and Information Systems, 52(1), 1-51.

[12] Hosseini, Z., &Jafarian, A. (2016). A Hybrid Algorithm based on Invasive Weed Optimization and Particle Swarm Optimization for Global Optimization. International Journal Of Advanced Computer Science And Applications, 7(10), 295-303.

[13] Nasim, R., &Kassler, A. J. (2017, May). A robust Tabu Search heuristic for VM consolidation under demand uncertainty in virtualized datacenters. In 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID) (pp. 170-180). IEEE.

[14] Zhang, L., Gao, L., & Li, X. (2013). A hybrid genetic algorithm and tabu search for a multi-objective dynamic job shop scheduling problem. International Journal of Production Research, 51(12), 3516-3531.

[15] Li, X., & Gao, L. (2016). An effective hybrid genetic algorithm and tabu search for flexible job shop scheduling problem. International Journal of Production Economics, 174, 93-110.

[16] Ren, Z., Chen, W., Zhang, A., & Zhang, C. (2013, July). Enhancing invasive weed optimization with taboo strategy. In Proceedings of the 15th annual conference companion on Genetic and evolutionary computation (pp. 1659-1662). ACM.

[17] Gopi, A., et al. "Designing an Adversarial Model Against Reactive and Proactive Routing Protocols in MANETS: A Comparative Performance Study." International Journal of Electrical & Computer Engineering (2088-8708) 5.5 (2015).

[18] Kumar, S. Ashok, et al. "An Empirical Critique of On-Demand Routing Protocols against Rushing Attack in MANET." International Journal of Electrical and Computer Engineering5.5 (2015).

[19] G. L. Sravanthi, M.Vasumathi Devi, K.Satya Sandeep, A.Naresh and A.Peda Gopi, “An Efficient Classifier using Machine Learning Technique for Individual Action Identification” International Journal of Advanced Computer Science and Applications(IJACSA), 11(6), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0110664.

Adaptive Retrieval Time-Related Data Model for Tracking Factors Affecting Diabetes

Ibrahim AlBidewi[1], Fahad Alotaibi[3]
Department of Information System
Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

Nashwan Alromema[2]
Department of Computer Science
Faculty of Computing and Information Technology-Rabigh
King Abdulaziz University, Rabigh, Saudi Arabia

***Abstract*—In the last four decades several dozens of representing time-oriented data/knowledge bases have been presented. Some of these representations violate First Normal Form (1NF) by using Non-First Normal Form (N1NF) prototypes and temporal nested representations, while others simulated the concepts of temporal data with relational data representation without violating 1NF. In this article, a new interval-based knowledge representational data model with an optimized retrieval techniques are employed for modeling and optimality retrieve a biomedical time-varying data (factors/observations that affect the diabetes). The used time-related data model is more compact to represent time-varying data with less memory (capacity) storage with respect to the main representations in the literature, but which is as expressive as those representations (a transformation algorithms show that data represented in this model can be transferred to/from the representations in the literature with zero percent loss of information). A new data structure is defined with the optimal retrieval techniques to prove some basic properties of the used time-model and to ensure that the time-model is an extension and reduction of the main representations in the literature, namely *TQuel* and BCDM. The expressive power, reducibility, and easy implementation of the proposed model, especially for the legacy systems, are considered as advantages of the proposed model.**

Keywords*—Diabetes database; time-data model; diabetes observations; valid-time data; knowledge-based data

I. Introduction

The reproduction method of Time-varying data representation in RDBMS is considered as bases of time-oriented information, and it provides temporal information models and stores data associated with the past, present and future. TDB also offers communicative and proficient ways to reproduce, stock up, and inquire about special temporal situation of the stocked information in contrast to the conventional databases which record single state of the real world phenomena. TDB is a vital area of study, with a dynamic population of a number of hundred investigators who have brought about thousands of research articles over the recent thirty years [1, 2, 4]. Conventional information record is utilized to stock and treat the information that refers to the present moment in time, without supporting the temporal features which maintain temporal database and store data referring to what went before, to current time and to upcoming time. There has been a discussion within the recent thirty years on how to represent, execute and inquiry about time-based information record in an efficient way. An increasing concern with time-varying information records in several practice fields which address matters in managing temporal information [2, 3, 4, 9, 14, 18, 26, 33]. Most of these publications touched upon the various features of time-varying information records.

The remaining part of this article is structured as it follows: Section 2 introduces the problem background and the existing techniques that deal with modeling time-varying data, Section 3 introduces the classification of time-varying information prototype expansions. Section 4 describes the representational methodology of time-data models. Section 5 compares three TDB prototypes in terms of memory storage representation. Lastly, Section 6 draws a conclusion and a close this article.

II. Background

Handling time-varying database can be accomplished by two approaches as mentioned in [5, 6, 10], namely, (1) The integrated approach is to build a complete TDB Management System from zero, which offers a primal information kind and deals with the various situations/time examples of the information being stocked. This approach involves bottom-up construction an entire TDBMS, which is a very huge and time-wasting job. It is also not easy since the fundamental values used by profitable DBMS to increase effectiveness of functions have to be improved and many of abstract works need to be performed to prove that the novel structure is entirely wholly perfect. The scope of time and workforce needed for this advance is parallel to that required by profit-making sellers to build up DBMS that all of us know nowadays. (2) The stratum method is to develop a strategy that expands time-independent information representation to time-varying information representation on top of common DBMS that behaves as a layer. The second method does not entail any alterations to the available information record technique. It can be merely elaborated by constructing an innovative method for time-varying backup in addition to the available common DBMS.

Modeling temporal database in relational framework by appending extra vertical divisions for time to the correlation(s) appears to be a straightforward way [5, 11, 12, 13, 20], yet it does not solve many delicate matters particularly for time-varying information. Many expansions of relational information representation to time-varying information

This project was funded by the Deanship of Scientific Research (DSR) under grant no. G- 27-005-1430.

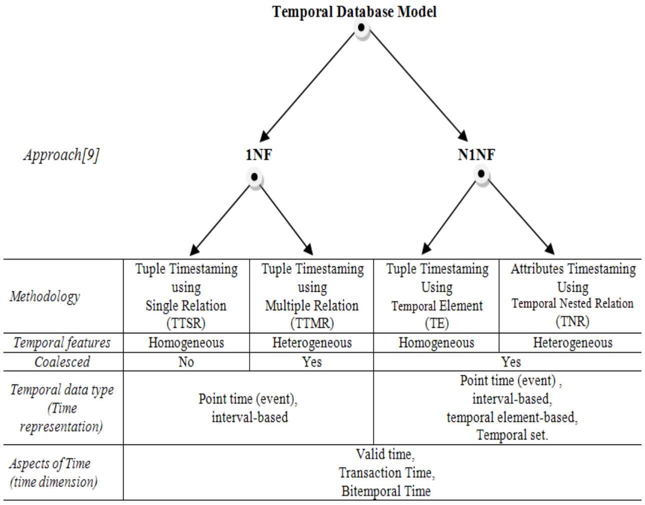
The authors, therefore, acknowledge with thanks to DSR's technical and financial support representations Methodology.

representation have been offered with their strengths and weaknesses [5]. More than 23 time-varying information prototypes have been proposed, some of these prototypes are in [13, 16-22, 27, 28,29-30].

In this paper, a taxonomy of all the possible extensions of Temporal Database Model (TDM) are developed as shown in Fig. 1 and according the naming conventions appear in the consensus glossary in [15]. Based on that, a novel information reproduction, a semantic time-varying relational data model that extends model in [33] (Snodgrass' Tuple Timestamped data representation) are introduced to deal with interval-based knowledge representation in relational data model. A new data structures, a cost model for the memory storage use for the proposed time-data model are defined. The extended model, and time-based relational algebra also provided to prove some essential characteristics and to ensure that the proposed model is an extension and reducible of BCDM.

III. Modeling Time-Varying Data

Modeling time-varying information in relational framework differs in many perspectives [9], the most recurrently declared methods are tuple timestamping with 1NF, and attribute timestamping with N1NF as shown in Fig. 1. Integrating time in relational data representation could be done by one of the TDB prototypes approaches that are shown in Fig. 1, which shows the approaches of modelling time-varying data in relational framework. These approaches can be either by applying 1NF or by violating 1NF using multi-value attributes or nested relations. The first method -1NF- has two differentiations, specifically; (1) Tuple Timestamping Single Relation (TTSR), this approach incorporates time in relational data model by adding extra timestamps attributes to the conventional association. Several time-varying information representation debated in [9] can be classified under this method. An instance of some of these time-varying information prototypes are LEGOL 2.0 by Jones [29], Temporally Oriented Data Model by Ariav [31], HSQL by Sarda [32], and TQuel by Snodgrass [33]. The time-varying information record representations can be harmonized or diverse. The time-varying harmonized feature of temporal information record relations [32] is defined as time-varying associations in which the temporal time of all information merits (that is to say the time over which they are determined) in each tuple is the same [9]. And as stated in [9, 24, 29] the representations that employ tuple timestamping are essentially temporally identical because only temporally uniform associations are probable, to conclude that the representations that are based on TTSR approach are temporally homogeneous. Temporally heterogeneous database relations are temporal database relations in which the temporal time of the information merits in each tuple can be dissimilar. Since 1NF tuple timestamped temporal relations that allow extension over or following tuples of equivalent values- are not coalesced [9] then the representations under this approach are not coalesced because tuples of equivalent values- are not unacceptable. The time representation of the stored facts can be event timestamp, or interval timestamps. Finally, this approach can model the diverse temporal time aspects like valid-time, transaction time, or Bitemporal time. TTSR method is not proficient since it presents tautology, where information merits that vary at diverse time are recurring in manifold tuples. (2) Tuple Timestamping Multiple Relations (TTMR), this method has resolved the issue of information tautology in TTSR by breaking down the temporal correlation as ensuing: Temporal values are spread over manifold associations, and non-temporal values are regrouped into a segregate association. Several time-varying information representations debated in literature [9] can be classified under this method. An instance of some of these time-varying information representations is the Temporal Relational Model by Navathe [30], and Snodgrass [8], and Tansel [16]. The temporal information representations under this method are attribute timestamping, thus, they are heterogeneous since the temporal time of the element worth in each tuple can be dissimilar, coalesced because value-equivalent tuples are disallowed. The time representation of the stored facts can be event timestamp, or interval timestamps. Finally, this approach can model the diverse time aspects like valid-time, transaction time, or Bitemporal time. The problem with the data representations under this approach is when the data for an entity is need to be to combined, a variant of bond known as temporal intersection join would be required, which is by and large costly to be accomplished. The second approach (N1NF) employs multi-value attributes, or in other words nested relations for tuple timestamping or attributes timestamping. As is shown in Fig. 1 the tuple timestamping approach uses sets of time chronons (the smallest time unit) for timestamping associated tuples. BCDM by Jensen and Snodgrass [9] is an example of this model. Prototypes under this category are homogeneous and coalesced. In attributes timestamping approach non-atomic attribute values are associated with temporal time. Historical Relational Data Model by Clifford-2 [33], HQuel by Tansel [34], Homogeneous Relational Model by Gadia-1 [32], Heterogeneous Relational Model by Gadia-2 [35], and TempSQL by Gadia-3 [35] are temporal data prototypes that can be considered under this category. N1NF may not be competent of directly utilizing a relational stocking construction or a questioning appraisal method that depends on tiny element merits, thus this approach (N1NF) will be excluded from this study. A variety of methods for managing time-varying information can be spotted in [1]. The traditional modelling technique for temporal database is not efficient since the prototypes that are based on 1NF model introduce redundancy, while the prototypes which are based on N1NF model may not be capable of directly implemented in conventional DBMS [9]. Using the integrated approach for implementing TDB prototypes in conventional DBMS is a costly task and can be accomplished only by the DBMS vendors [6]. Therefore, many temporal database prototypes have been anticipated in attempting to seize the time-varying attribute of information in ease of use and system performance [2, 34, 27]. The implementation of TDB in conventional DBMSs is not an easy task [34], this is because conventional DBMSs do not offer prototypes to back up and treat the temporal dimensions of the implemented database [5, 11]. Because of the unnecessary tautology of information, costly execution, and the complexity of implementing TDB in relational framework, an intelligent way is required to represent, execute, and question TDB in correlation outline.



Methodology	Tuple Timestaming using Single Relation (TTSR)	Tuple Timestaming using Multiple Relation (TTMR)	Tuple Timestaming Using Temporal Element (TE)	Attributes Timestaming Using Temporal Nested Relation (TNR)
Temporal features	Homogeneous	Heterogeneous	Homogeneous	Heterogeneous
Coalesced	No	Yes	Yes	
Temporal data type (Time representation)	Point time (event), interval-based		Point time (event) , interval-based, temporal element-based, Temporal set.	
Aspects of Time (time dimension)	Valid time, Transaction Time, Bitemporal Time			

Fig. 1. The Taxonomy of Temporal Data Model.

This section will begin by presenting the proposed TDB model for modelling the biomedical data. This data model is termed as Tuple Timestamp Historical Relation (TTHR). The generic means of temporal database is an information record with some sort of temporal back up. In this paper the focal point is on the valid time characteristic of time-varying information in relational information records; nevertheless, the debate might simply be expanded to information records that sustain transaction-time and bitemporal time. Then the semantics of associating time aspect to temporal database objects as well as representing these objects in TTHR will be explained in this section. Time Domain is assumed linear, bounded, and discrete constructional features of the temporal field for timestamping the proposed TDB representation as in [9]. Chronons (the smallest time unit) are utilized that have length and limited precision [14]. The temporal domains have overall orders and are similar in form and relations to the domain of ordinary figures. Associating Time Aspects to the Database Objects can be affirmed as valid time and transaction time where together are conceived as the universal features of all information record data [28, 29]. Since the data in the information schemes are built by a number of characteristics, then valid time and transaction time characteristic related to a group of aspects might be employed to record the valid time of the data, or the time when the data are present in the information record. The duration time and transaction time are universal features of the unit types or association types. The lifetime of an information record entity as in [29] is the moment in time over which it is identified, or the moment in time when the matching entity exists in the reproduced reality. The valid-time lifespan of an information record object refers to the point in time when the entity exists in the constructed reality, whereas the transaction-time lifespan of an information record object refers to the point in time when the object is present in the information record. The information record object's lifespan can be reduced to valid-time as mentioned in [31]. In this study, only the valid-time lifetime of an information record entity is considered, where the matching entity can have one or more time-varying attributes. For these time-varying attributes, a valid-time aspects is associated in such a way that the valid-time aspects are rigorously enclosed in or equal to the lifespan of the object (example valid-time interval [3, 7] for lifespan interval [1, 34], for short the term lifespan time aspect is used instead of valid-time lifespan to avoid the confusion). In general, an object type in an abstract data representation which matches with one relational schema in the rational data representation can be associated with the subsequent time features: (1) Lifetime period feature, if the lifetime of that object sort is a division of the represented moment in time, the object category will be associated with the lifetime period features. (2) Transaction time aspect indicates the presented time of the object in the information record. Association categories in an abstract representation which matches with one relational schema in the rational representation can be attributed by: (1) Lifetime period feature where the lifetime of the association kind should be a division of the lifetime of the contributing objects. (2) Transaction time, where the meaning of the transaction time features of the association kind is a division of the lifetime or transaction-time features of the contributing objects. The time-varying backup for association kinds (encompassing lifetime and transaction-time) may be as an immediate occurrence where the characteristics of these association kinds are thought of as predetermined (time-independent) features. Also, it may be a lasting occurrence where the characteristics of this association kind may be temporal features, or the non-temporal representation of these characteristics. The characteristics in the abstract representation that are matching with the vertical divisions of a chart in the rational representation might be regarded as temporal or non-temporal characteristics. In case of time-varying attributes, the time aspects that can be assigned to these attributes are as follows: (1) Valid-time feature, where these time features are supposed to be a division of the lifetime of the related object kinds or association sort. (2) Transaction-time, where these temporal features are supposed to be a division of transaction-time of the related object kind or association sort. (3) Bi-temporal time features, where these temporal features have the composite constraints of valid-time and transaction-time features. This temporal relationship and the constraint of characteristics are applicable for the characteristics that ensue from single-to-multiple association. The data in Table I depicts the features of time for which they ought to be supplied for the information record items and as shown in [12]. The temporal features that might be connected to diverse information record items are dissimilar from one information record item to another. For instance, a lifetime period may not be connected with characteristics as it shows the availability of the item or unit sort in the modeled reality. Thus, the relationship of a point in time with a diverse information record item ought to be constrained in accordance with Table I.

TABLE I. INFORMATION RECORD ITEMS AND THEIR PROPPED FEATURES OF TIME

Database object	Lifespan (LS)	Valid Time (VT)	Transaction Time (TT)	Bi-temporal (Bi)
Entity (E)	Yes	No	Yes	No
Attribute (A)	No	Yes	Yes	Yes
Relationship (R)	Yes	Yes	Yes	Yes

Coding that, a lifetime moment in time can be coupled with a correlation kind that has been being by itself; however, in case the association is carried out as an aspect of one of the constituting objects, it will be processed as an aspect and cannot be coupled with a lifetime scope.

A. Representing and RetrievingTime-Vaying Data (Diabetes Data) using TTHR Approch

A temporal object in *TTHR* can be associated with and restricted to time aspects stated in Table I. The valid time features for attribute objects, and lifetime aspects for entity and relationship objects are considered as a timestamping method in this article (it could trivially extend this model by replacing lifespan time aspects by transaction-time aspects, with such an expansion, the proposed model is insignificantly a constant expansion of BCDM [9, 29]). The field of applicable times can be represented as $D_{VT} = \{t_1, t_2, \cdots, t_{VT}\} \cup \{\infty\}$ and the field of lifetime intervals can be represented as $D_{LS} = \{t_1, t_2, \cdots, t_{LS}\} \cup \{\infty\}$, where ∞ is a *distinguished value* that represents a very large value for time (say 3000 ac). The time domain of both *VT* and *LS* is an order of natural numbers. In what follows, R_T is employed to indicate the schema of temporal relation. The group of attributes, which constitute R_T may be classified into ***k***ey attributes ***(K)***, non-temporal attributes (***U***nchangeable ***(U)***), temporal attributes (***C***hangeable***(C)***), and ***T***imestamps attributes (***T***). $R_T = \{A_K, A_U, A_C, A_T\}$ .Where $A_K = \{A_{K1}, A_{K2}, \cdots, A_{Ki}\}$, $A_U = \{A_{U1}, A_{U2}, \cdots, A_{Un}\}$, $A_C = \{A_{C1}, A_{C2}, \cdots, A_{Ci}\}$, and $A_T = \{T_{ls}, T_{le}\}$, the domain of T_{ls} and T_{le} is D_{LS} , where T_{ls} represents the Lifespan Start Time (*LSST*) and T_{le} represents the Lifespan End Time (*LSET*). Temporal relational schema R_T in *TTHR* can be represented as $R_T = \{A_K, A_U, A_C, A_T\}$, and a new secondary association representation VT_R_T is formed as $VT\_R_T = \{A_K, index, \alpha, A_T\}$ where *index* is a factor to spot the time-varying aspect A_{Cm} begin updated (where $1 \le m \le i$), α is the updated value of A_{Cm} and $A_T = \{T_{vs}, T_{ve}\}$, the domain of T_{vs} and T_{ve} is D_{VT} , where T_{vs} constitutes the Valid Start Time (*VST*) and T_{ve} constitutes the Valid End Time (*VET*). The aim of this modeling is to maintain the most recent (existing) updated information in one relation R_T , and the chronological alterations of the logical soundness of the temporal information in a secondary association VT_R_T . An example of correlation is indicated by r_t , and vt_r_t , where $r_t(R_T)$ implies r_t is an example of R_T , and $vt\_r_t(VT\_R_T)$ implies vt_r_t is an example of VT_R_T . For tuples x, y and z are employed**,** accordingly a tuple, $x = (a_K, a_U, a_C, a_T)$ in the association example $r(R_T)$ is composed of a number of attribute values associated with the T_{ls} and T_{le} , whilst the tuple(s) $vt\_x = (a_K, index, \alpha, a_T)$ in the relation instance $vt\_r_t(VT\_R_T)$ is/are strongly correlated and in reference to tuple x **.** The tuple(s) is composed of the principal key of x , the characteristics (*index*) of the temporal aspect in x , the updated temporal features value α in x , and the validity of the updated characteristic T_{vs} and T_{ve} . A division of the field of natural life time is connected to each tuple in R_T demonstrates that the availability of the item registered by the tuple is factual in the reproduced veracity all through every lifetime chronon in that division. A division of field of valid times is coupled with every tuple in VT_R_T , denotes the reality that the tuple vt_x registers the alteration of the validity of a_{cm} in x **.** This reality is regarded as factual in the reproduced veracity throughout each valid time chronon in that division, noting that the time of validity is stringently comprised in the temporal lifetime of x . Therefore, the coupled time with a tuple in *TTHR* is interval-based temporal timestamp. The semantics of the proposed data model can be informally explained via a simple example as show in Fig. 2, where a temporal relation schema *Patients* corresponding to R_T in *TTHR* is used to record *Patients* information. The auxiliary relation *VT_Patients* is employed to register the alterations of the validity of the time-varying characteristics in *Patients* and the alterations of the lifetime of the items in *Patients*. The diverse kinds of characteristics of *Patients* and *VT_Patients* are:

$$A_K = \{p\_id\}$$

$$A_U = \{p\_name, Birth\_date, sex, family\_\inf ected\}$$

$$A_C = \{weight, hight, BMI, GF, Cho, Tri, Sua, BP, Hyp,$$

$$GTT, HBLIC, Fru, RBG, PBG, GTT, HBLIC$$
$$Fru, RBG, PBG\}$$

$$A_T = \{T_{ls}, T_{le}\}$$

Semantically, the attributes of the Patients relation have the following meaning: P_Id– Patient's identified number, Name- Patient's name, Birth_date- Patient's birth date, sex- Patient's gender, Family_infected- a Boolean attributes to answer either any of Patient's relative had sugar or not, nationality- Patient's ethnicity, Weight- Patient's weight, Height- Patient's height, BMI- Patient's Body mass index. The rest of the attributes are sugar related test observations, the glossary of these terms and the meaning of each can be found at "Joslin Diabetes Center". Table II outlines the full observations' names, descriptions, the test range, and the abbreviation of each one. The temporal database relation *Patients_VT* (Fig. 3) is an auxiliary relation of patient's relation that recodes the tracking of different diabetes observations changes of the patients. It consists of four attributes, *P_id, Att_index, Updated_V,* and temporal attributes T_{vs} and T_{ve} , such that T_{vs} represents the Valid Start Time (*VST*) and T_{ve} represents the Valid End Time (*VET).* The semantics of *patients_VT'* attributes are, *P_id*: is a foreign key referencing to *P_id* in patients relation. *Att_index*: contains data about the locations of time-varying attributes in patient's relation. *Updated_V*: contains the old data (observations' result) of the updated time-varying attributes in patients relation that is indexed by *Att_index*. T_{vs} : is a temporal attribute that represents (*VST*), andT_{ve} : is a temporal attribute that represents (*VET*).

TABLE II. LIST OF OBSERVATIONS FOR DIABETES PATIENTS' TESTS [36, 37]

Observation	Test	
	Descriptions	Normal Range
Glucose Fasting (GF)	a blood test in which a sample of blood is drawn after an overnight fast to measure the amount of glucose in blood.	65-110 mg/dL
Cholesterol (Cho)	Cholesterol is tested at more frequent intervals	up to 200 mg/dL
Triglycerides (Tri)	This tests is used to help identify an individual's risk of developing heart disease.	up to 150 mg/dL
Serum Uric Acid (SUA)	This test is used to diagnose gout.	2.50-6.50 mg/dL
Blood pressure(BP)	the pressure of circulating blood on the walls of blood vessels.	120/80- 140/90
Hypertension (Hyp)	is defined as blood pressure higher than 140 over 90 mmHg (millimeters of mercury). .	90-140 and above
Glucose Tolerance Test(GTT)	A glucose tolerance test measures how well your body's cells are able to absorb glucose, or sugar, after you ingest a given amount of sugar.	95-180 mg/dL
Glycosylated Haemoglobin (HBA1c)	A form of hemoglobin that is measured primarily to identify the three-month average plasma glucose concentration.	4.5-7
Fructosamine (Fru)	Fructosamines are compounds that result from glycation reactions between a sugar.	200-285 µmol/L
Random Blood Glucose (RBG)	Random blood sugar (RBS) measures blood glucose.	80-140 mg/dL
Postprandial Blood Glucose (PBG)	A postprandial glucose test is a blood glucose test that determines the amount of a type of sugar, called glucose, in the blood after a meal.	110-160 mg/dL
Microalbumin Test (MicT)	A test to detect very small levels of a blood protein (albumin) in your urine.	< 30 mg of protein
Serum Creatinine Test (SCT)	A creatinine blood test measures the level of creatinine in the blood.	0.5 to 1.1 mg/dL (female) 0.6 to 1.2 mg/dL (male).
Low Density Lipoproteins(LDL)	LDL is called low-density lipoprotein because LDL particles tend to be less dense than other kinds of cholesterol particles.	Up to 130 mg/dL
High Density Lipoproteins(HDL)	HDL is known as the good cholesterol because it carries LDL, triglycerides, and harmful fats and returns them to your liver for processing.	> 35mg/dL
Uric Acid (UA)	Uric Acid is excreted (removed from your body) in your **urine**.	2.4-6.0 mg/dL (female) 3.4-7.0 mg/dL (male).

Patients

	1	2	3	4	5	
P_ID	*P_Name*	*Birth_date*	*Sex*	*Family*	*Nationality*	
101	Jon	1/10/1990	M	Y	UK	…..
108	Sara	1/10/1990	F	N	MY	…
109	Sonya	1/10/1975	F	Y	YE	…
102	Tony	1/5/1986	M	Y	SA	…

………

6	7	8	9	10	11	…
Weight	*Height*	*BMI*	*GF*	*Cho*	*Tri*	…..
70	160	N	170	230	211	…
80	166	N	150	100	139	…
75	155	N	367	203	207	…
80	169	N	115	180	275	….

… 12	13	14	15	16	17	18	19
SUA	*BP*	*Hyp*	*GTT*	*HBA1c*	*Fru*	*RBG*	*PBG*
6.3	70	99	98	4.6	204	83	130
5.7	80	96	103	4.7	208	90	100
2.8	90	103	109	6	250	85	130
6.3	100	134	150	5.2	230	106	150

………………..

20	21	22	23	24		
MicT	*SCT*	*LDL*	*HDL*	*UA*	T_{ls}	T_{le}
20	0.7	140	50	4	20	∞
25	0.8	150	90	4.5	10	∞
29	1	300	100	5.3	19	∞
17	0.9	200	140	6	50	∞

Fig. 2. The Representation of Patients Database relation using TTHR Approach.

Patient_VT

P_ID	*Att_index*	*Updated_V*	T_{vs}	T_{ve}
102	9	108	50	67
102	10	196	50	67
102	11	381	50	67
102	9	102	68	90
102	10	209	68	90
102	11	284	68	90
102	12	6.58	50	90
109	9	241	19	105
109	8	O	19	100
108	9	300	10	110
108	9	290	111	150
108	9	250	151	160
108	9	210	161	170
108	9	190	171	190
102	13	109	50	70
102	14	126	50	90
102	19	140	50	85
102	23	118	50	98
109	13	100	19	97
109	14	112	19	77

Fig. 3. Auxiliary Temporal Database relation of Diabetes Patients.

The methodology of the retrieval techniques of the proposed time-varying data model is different from conventional databases because temporal database holds a sequence of snapshot relations. Therefore, there are three different types of queries, namely, current query, sequenced query and non-sequenced query [8, 20]. The most prevalent queries (current) are of the form "what is valid now?", the second type of queries (non-sequenced) is of the form "what was valid at any time?", and the third type of queries (sequenced) are of the form "what is valid at/during a certain point/interval (period) of time?". The valid time data represented by the proposed data model will be shown how it can be queried using the different types of temporal query. For querying the biomedical data, an interested queries according to the diabetes medical doctor is asking about the latest reading of any of the 19s observations as in Fig. 2, the reading values of any observations in any point of time, and the reading values of any observations when the values of other observations has specific reading. Queries B1 to B2 with Fig. 4, Fig. 5, and Fig. 6 depict and show the results of such queries.

Query B$_1$: What is the latest reading of *Glucose Fasting* of Tony with *P_ID =102*?

$$Q_{B_1} \leftarrow \pi_{\text{P\_Id,P\_}Name\text{,Birth\_date,GF}}(\sigma_{P\_Id=102}(Patients))$$

P_ID	*P_Name*	*Birth_date*	*GF*
102	Tony	1/5/1986	115

Fig. 4. The Result of Glucose Fasting Query.

Query B$_2$: What are the readings of *Glucose Fasting* of Tony with *P_ID =102* during the period of [60, 100]?

$$Q_{B_2} \leftarrow \sigma_{\substack{T_{vs} \le 100 \\ and \\ T_{ve} \ge 60}}(P\_\text{GF})$$

P_ID	*P_Name*	*GF*	T_{vs}	T_{ve}
102	Tony	108	60	67
102	Tony	102	68	90
102	Tony	115	91	100

Fig. 5. The Result of Glucose Fasting Query in the Period [60, 100].

Query B$_2$: What are the readings of *Glucose Fasting* of Tony with *P_ID =102* when the reading of *Serum Uric Acid* was not in the normal range?

For this query, the period of validity of *Serum Uric Acid* when it was in up normal reading (according to Table II this observation value ranges from 2.5 to 6.5) will be retrieved according to the query Q_{SAU} below. The overlap period of validity of *Glucose Fasting* with the period of validity of *Serum Uric Acid* when it was in up normal reading will give the result as shown in Fig. 4 to 6.

$$Q_{SUA} \leftarrow \sigma_{P\_Id \le 102..and..SUA > 6.5}(P\_SUA)$$

P_ID	*P_Name*	*GF*	T_{vs}	T_{ve}
102	Tony	108	60	67
102	Tony	102	68	90
102	Tony	115	91	100

SAU_Upnormal

$$Q_{B_3} \leftarrow ((\text{P\_GF} \underset{\substack{GF\_P\_id = SAU\_P\_id \\ and \\ \text{GF\_}T_{vs} \le SAU\_T_{ve} \\ and \\ \text{GF\_}T_{ve} \ge SAU\_T_{vs}}}{\rhd\lhd} SAU\_Upnormal))$$

P_ID	*P_Name*	*GF*	*SAU*	T_{vs}	T_{ve}
102	Tony	108	6.58	60	67
102	Tony	102	6.58	68	90

Fig. 6. The Result of Glucose Fasting Query when Serum Uric Acid was up Normal.

B. Representing Time-Vaying Data using TTSR Approchs

More than 23 time-varying data models have been introduced in the literature [9]. These data prototypes are categorized and compared with respects to fundamental design decisions that are represented by asking how valid time is modeled, how transaction time is modeled, how attribute values are modeled, whether the model is all the same, and whether the model stick together as a whole. These five questions represent the criteria of evaluating temporal data model according to TSQL2 [9]. To generalize the approaches of modeling time-varying database, the taxonomy shown in Fig. 1 depicts that. *TTHR* is a combination of tuple timestamping and attribute timestamping data prototypes. Whereby the facts are timestamped by the lifespan of the associated entities, and the time varying attributes of these entities are timestamped by the valid time as explained in the previous section. *TTHR* is 1NF, heterogeneous, and interval-based temporal data model which is dedicated for modeling temporal database in relational framework. In this part, the modeling of *TTHR* in *TTSR* is scrutinized. For each approach/model, we stipulate the items identified in the model, present the associations to/from *TTHR* model to show that the same data are being stocked without losing any information. The study is progresses from the different approaches that are based on 1NF prototypes, and exclude N1NF prototypes which are not in the scope of this work. *TTSR* and *TTMR* approaches shown admit different representations, this work chose the representations explained in [18, 2] which are more close to the used time-model and easy to be implemented in conventional DBMS for the purpose of comparative analysis study. In *TTSR* the associations are reproduced by snapshot relations, which are 1NF associations. In below an illustration on how to embody *TTHR* associations in the *TTSR* representation is shown. Let R_{TTSR} stand for the association in the *TTSR* model that has the diagram representation $R_{TTSR} = (A_K, A_U, A_C, T_{ls}, T_{le}, T_{vs}, T_{ve})$ the corresponding relations in the *TTHR* representation are $R_T = (A_K, A_U, A_C, T_{ls}, T_{le})$ for current valid data and $VT\_R_T = (A_K, index, \alpha, T_{vs}, T_{ve})$ for historical changes of time-varying data. Since the information prototypes characterize associations in a different way and, for clearness and regularity of details, a correlation examples in tabular form is provide.

C. Representing Time-Vaying Data using TTMR Approaches

In *TTMR* the associations are accounted for by snap relation $R_{TTMR} = (A_K, A_U, T_{ls}, T_{le})$, for each temporal characteristic there are distinct associations $R_{A_{c1}} = (A_K, A_{C1}, T_{vs}, T_{ve}) \ldots R_{A_{ci}} = (A_K, A_{Ci}, T_{vs}, T_{ve})$, and for the lifespan time $R_{LS} = (A_K, T_{ls}, T_{le})$, which are all in 1NF relations. Below describes how to represent *TTHR* relations in *TTMR* representation. The corresponding relations in *TTHR* representation are $R_T = (A_K, A_U, A_C, T_{ls}, T_{le})$ for current valid data and $VT\_R_T = (A_K, index, \alpha, T_{vs}, T_{ve})$ for historical changes of time-varying data. Since the information prototypes represent associations in a different way and, for both transparency and constancy of details the relation examples is shown in tabular form. A time-varying representation *patients* in Fig. 2, parallel to R_{TTMR} in *TTSR* is divided into $(i+2)$ associations, where *i* is equal to 19 (number of time-varying attributes) and the 2 other associations are the lifetime association, and the association that entails the non-temporal characteristics. The 19 associations parallel to each temporal value will be employed to register the historical changes of the validity of the temporal characteristics in *patients*. The lifetime correlation will be employed to trace the alterations of the lifetime of the items in *patients* relation and ultimately the non-time varying association is employed to register the non-temporal characteristics.

IV. Comparison of Temporal Data Models

This part will match up the three methods/prototypes in terms of information illustration and memory storage evaluation, a similar work done in [23] using different parameters. The start will be with the evaluation in terms of memory storage Point of View, Let *R* be a time-varying associational representation with a set of attributes $\{A_1, A_2, \cdots, A_n, T\}$, where these attributes can be categorized into 4 categories: ***k***ey, non-temporal (***U***nchangeable), temporal (***C***hangeable), and ***T***imestamps. They are represented by ***K***, ***U***, ***C***, and ***T*** correspondingly. Consequently, the representation of time-varying association can be re-represented as $\{A_K, A_U, A_C, A_T\}$, where $A_K = \{A_{K1}, A_{K2}, \cdots, A_{Kj}\}$, $A_U = \{A_{u1}, A_{u2}, \cdots, A_{un}\}$, $A_C = \{A_{c1}, A_{c2}, \cdots, A_{ci}\}$ and $A_T = \{A_{T1}, A_{T2}\}$ The timestamp attributes are defined as $A_{T1} = T^{c1}$, $A_{T2} = T^{c2}$. The subscripts variables *j, n*, and *i* represent the total number of key attributes (A_K), total number of Time-invariant attributes (A_U) and total number of Time-varying attributes (A_C), respectively.

Definition 1 (Non-temporal characteristic), A characteristic the value of which does not alter with time, a Non-temporal characteristic can be reorganized as in the case of an inaccuracy, but temporal information record does not maintain a record of it.

Definition 2 (Temporal characteristic), A temporal characteristic is a feature the merits of which are connected with timestamps.

Definition 3 (Timestamp), A timestamp is a temporal worth correlated with a Timestamped item (i.e., a feature worth or tuple).

Definition 4 (Lifespan), the lifetime of record items is the moment in the course of which the item is outlined.

Definition 5 *(Frequency of Time-varying attribute* $f(A_{cm})$, the amount of occurrence rate this characteristic is to be altered (changed). Where $A_C = \{A_{c1}, A_{c2}, \cdots, A_{cm}, \cdots, A_{ci}\}$ and $m \in \{1, 2, \ldots, i\}$. Thus, a function f defined on A_C in such a way that $f(A_{cm})$ which returns the frequency of times A_{cm} will be changing.

Definition 6 ($S(A_{\chi\gamma})$), A role to be determined on all the characteristics of R, where $S(A_{\chi\gamma})$ reverts the magnitude of a characteristic $A_{\chi\gamma}$ in bytes, and $\chi \in \{K, U, C, T\}$ and $\gamma \in \{1, 2, \ldots, j\}$ (key attributes), or $\gamma \in \{1, 2, \ldots, n\}$ (non-temporal features) or $\gamma \in \{1, 2, \ldots, i\}$ (temporal characteristics) or $\gamma \in \{1, 2\}$ (timestamping traits) for all attributes sets that construct R.

Definition 7 $Cost(A_\chi)$, A role to be determined on the subclass characteristics χ, where $\chi \in \{K, U, C, T\}$. $Cost(A_\chi)$ reverts the total of all characteristics amount in A_χ in bytes.

Definition 8 $Cost(z)$, the cost of a tuple (line) z in relation example r_t is the total of the cost of all sub-group characteristics equivalent to $\operatorname{cost}(A_K) + \operatorname{cost}(A_U) + \operatorname{cost}(A_C) + \operatorname{cost}(A_T)$.

Definitions 9 The cost of a different attribute type is defined as:

$$\operatorname{cost}(A_k) = \sum\nolimits_{i=1}^{j} \operatorname{cos}t(A_{ki}) = K\ byte \tag{1}$$

$$\operatorname{cost}(A_u) = \sum\nolimits_{i=1}^{n} \operatorname{cos}t(A_{ui}) = U\ byte \tag{2}$$

$$\operatorname{cost}(A_c) = \sum\nolimits_{m=1}^{i} \operatorname{cos}t(A_{cm}) = C\ byte \tag{3}$$

$$\operatorname{cost}(A_T) = \sum\nolimits_{i=1}^{2} \operatorname{cos}t(A_{Ti}) = T\ byte \tag{5}$$

Definitions 10 The Frequency time of all time varying attributes $A_{cm} \in \{A_{c1}, A_{c2}, \cdots, A_{ci}\}$ in a time span λ can be computed as:

$$f(A_c) = \sum\nolimits_{m=1}^{i} f(A_{cm}) = \delta\ times \tag{5}$$

The evaluation of the *TTHR* in terms of memory storage used for physical implementation in conventional DBMS should encounter the same issues for the other time-varying information prototype. Since representing time-varying information record in a co-relational structure varies in numerous proportions [9]. Since the second approach (N1NF) may not be able of straightforwardly employing a correlation stocking structure or a uncertainty appraisal technique which depends on minute characteristic merits [9], then the temporal database prototypes that use this approach will not be counted in this comparison. The storage point of view in the *TTHR* will be compared with temporal database prototypes that use tuple timestamping with 1NF. In contrast with the employed storage in diverse prototypes, a predetermined time-span not across records for the information record file construction model was applied in this study. Assume $R = \{A_K, A_U, A_C, A_T\}$ as defined previously, we then use the different approaches and compare them with the *TTHR* model, a similar study has been carried out in [7].

For TTSR model, the time-varying association in *TTSR* can be modeled as:

R_{TTSR} ($A_{k1}, \cdots, A_{ki}$, $A_{u1}, \cdots, A_{un}$, $A_{c1}, \cdots, A_{cm}$, A_T), The cost of modeling one tuple x in the association example $r(R_{TTSR})$ can be computed as:

$$\begin{aligned}&\text{Cost}(x) = \text{cost}(A_k) + \text{cost}(A_u) + \text{cost}(A_c) + \text{cost}(A_T)\\&= K + U + C + T \ \mathit{byte}\end{aligned},$$

as indicated in (1), (2), (3) and (4).

The cost of stocking the record of the alterations of A_c with $f(A_c) = \delta$ frequency in a time span λ can be computed as:

$$= \delta(K + U + C + T) \tag{6}$$

A change in any A_c needs the inclusion of a new line with the entire values.

In the *TTHR* prototype, the time-varying association scheme is accounted for by R_{TTHR} and R_{VT} as indicated underneath:

R_{TTHR} ($A_{k1}, \cdots, A_{ki}$, $A_{u1}, \cdots, A_{un}$, $A_{c1}, \cdots, A_{cm}$, A_T), R_{VT} ($A_{k1}, \cdots, A_{ki}$, $Index, \alpha$, A_T). The cost of modeling one tuple x in relation example $r(R_{TTHR})$ can be computed as:

$$\begin{aligned}&\text{Cost}(x) = \text{cost}(A_k) + \text{cost}(A_u) + \text{cost}(A_c) + \text{cost}(A_T)\\&= K + U + C + T \ \mathit{byte}\end{aligned} \tag{7}$$

As indicated in (1), (2), (3) and (4). The cost of stocking up the record of the alterations of A_c with $f(A_c) = \delta$ occurrences in a time span λ can be computed as:

$$\begin{aligned}&= \delta(K + index + \alpha + T) \Rightarrow\\&= \delta(K + 1 + \alpha + T)\end{aligned} \tag{8}$$

Index: is a new feature to list the temporal values with one byte range:

α : is a novel extra value of an alternative information category to seize information from a diverse type. Its range is supposed to be equal to the range of the biggest domain range in A_c . The range of α in byte is $S(\alpha) = Max(S(A_{c1}), S(A_{c2}), \cdots, S(A_{ci}))$. Since every change in any A_c needs introducing a novel line in R_{VT} for the previous worth of the brought about (altered) temporal values, the sum is stored in a memory bank (room) for *TTHR* over *TTSR* for a definite time span λ can be computed as is shown below:

From (6), $\text{Cost}(TTSR) = (K + U + C + T) + \delta(K + U + C + T)$

From (7) and (8),

$\text{Cost}(TTHR) = (K + U + C + T) + \delta(K + 1 + \alpha + T)$

Cost (improvement)

$$\begin{aligned}&= \frac{\text{Cost}(TTSR) - \text{Cost}(TTHR)}{\text{Cost}(TTSR)}\\&= \frac{K + U + C + T + \delta(K + U + C + T) - (K + U + C + T + \delta(K + 1 + \alpha + T))}{K + U + C + T + \delta(K + U + C + T)}\\&= \frac{\delta(K + U + C + T) - \delta(K + 1 + \alpha + T)}{(K + U + C + T)(1 + \delta)}\\&= \frac{\delta(U + C - 1 - \alpha)}{(K + U + C + T)(1 + \delta)}\end{aligned}$$

Since $\alpha >> 1$ then cost (improvement)

$$\approx \frac{U + C - \alpha}{K + U + C + T} \tag{9}$$

Assume $Q = U + C$ then (9) will be:

$$\approx \frac{Q - \alpha}{K + Q + T}$$

For TTMR approach, the spared space in the storeroom for the *TTHR* over the *TTSR* would be directly proportional to $Q = U + C$ for a very large value of Q comparing to $K, T, and \alpha$ since this is due to the repeated data in *TTSR*.

The time-varying association in *TTMR* is decomposed and reproduced as:

R_{TTMR} ($A_{k1}, \cdots, A_{kj}$, $A_{u1}, \cdots, A_{un}$)

R_{Ac1} ($A_{k1}, \cdots, A_{kj}$, A_{c1} , A_T)

R_{Ac2} ($A_{k1}, \cdots, A_{kj}$, A_{c2} , A_T)

R_{Ac3} ($A_{k1}, \cdots, A_{kj}$, A_{c3} , A_T)

R_{Ac4} ($A_{k1}, \cdots, A_{kj}$, A_{c4} , A_T)

$$R_{Aci}\ (A_{k1}, \cdots, A_{kj}\ , A_{ci}\ , A_T\ )$$

The rate of stocking one tuple x in an association example $r(R_{TTMR})$ can be computed as follows:

From (1), (2), (3) and (4),

$$\mathrm{Cost}(x) = \mathrm{Cost}(A_k) + \mathrm{Cost}(A_u) + \mathrm{Cost}(A_T)$$
$$+\sum_{m=1}^{i}\left[S(A_{cm}) + \mathrm{Cost}(A_k) + \mathrm{Cost}(A_T)\right]$$
$$= K + U + T + \sum_{m=1}^{i} S(A_{cm}) + \sum_{m=1}^{i} K + \sum_{m=1}^{i} T$$
$$= (K + U + T) + i(K + T) + \sum_{m=1}^{i} S(A_{cm})$$

Since

$$\mathrm{Cost}(A_c) = i(K+T) + \sum_{m=1}^{i} S(A_{cm})$$
$$= i(K+T) + C$$

Then the $\mathrm{Cost}(x)$ can be reproduced as:

$$\mathrm{Cost}(x) = K + U + T + C + i(K+T)$$
$$= K(i+1) + U + C + T(i+1) \quad \mathit{byte} \tag{10}$$

Where *i* is the total number of A_c .The rate of stocking the record of the updates of each A_{cm} with $f(A_{cm}) = \delta_m$ frequencies in a time span λ can be computed as:

$$= \sum_{m=1}^{i} \delta_m (S(A_{cm}) + K + T)$$
$$= \sum_{m=1}^{i} \delta_m S(A_{cm}) + (K+T)\sum_{m=1}^{i} \delta_m$$

Since form (5)

$$\sum_{m=1}^{i} \delta_m = \delta$$, then

$$= \sum_{m=1}^{i} \delta_m S(A_{cm}) + \delta(K+T) \tag{11}$$

The improvement of the *TTHR* over the *TTMR* in terms of storage point of view can be calculated as follows:

The cost of storing one tuple x with its history of the changes in a period of time λ using the *TTHR* and the *TTMR* can be calculated as:

From (7) and (8),

$$\mathrm{Cost}(TTHR) = (K + U + C + T) + \delta(K + 1 + \alpha + T)$$

From (10) and (11),

$$\mathrm{Cost}(TTMR) = K(i+1) + U + C + T(i+1) +$$
$$\sum_{m=1}^{i} \delta_m S(A_{cm}) + \delta(K+T)$$

Cost (improvement)

$$= \frac{\mathrm{Cost}(TTMR) - \mathrm{Cost}(TTHR)}{\mathrm{Cost}(TTMR)}$$

$$= \frac{i(K+T) + \sum_{m=1}^{i} \delta_m S(A_{cm}) - \delta(1+\alpha)}{iK + k + U + C + iT + T + \sum_{m=1}^{i} \delta_m S(A_{cm}) + \delta K + \delta T}$$

$$= \frac{iK + K + U + C + iT + T + \sum_{m=1}^{i} \delta_m S(A_{cm}) + \delta K + \delta T - K - U - C - T - \delta K - \delta - \delta\alpha - \delta T}{K(i+1) + U + C + T(i+1) + \sum_{m=1}^{i} \delta_m S(A_{cm}) + \delta(K+T)}$$

$$= \frac{iK + iT + \sum_{m=1}^{i} \delta_m S(A_{cm}) - \delta - \delta\alpha}{K(i+1) + U + C + T(i+1) + \sum_{m=1}^{i} \delta_m S(A_{cm}) + \delta(K+T)}$$

$$= \frac{i(K+T) + \sum_{m=1}^{i} \delta_m S(A_{cm}) - \delta(1+\alpha)}{K(i+1+\delta) + U + C + T(i+1+\delta) + \sum_{m=1}^{i} \delta_m S(A_{cm})}$$

Since $\alpha >> 1$ and $\delta >> 1$ then the equations will be

$$= \frac{i(K+T) + \sum_{m=1}^{i} \delta_m S(A_{cm}) - \delta\alpha}{K(i+\delta) + U + C + T(i+\delta) + \sum_{m=1}^{i} \delta_m S(A_{cm})} \tag{12}$$

Let $\omega = \sum_{m=1}^{i} \delta_m (S(A_{cm}))$ then $\omega \approx \delta\alpha$ when the size $S(A_{cm})$ of all $A_{cm} \in \{A_{c1}, A_{c2}, \cdots, A_{ci}\}$ is almost the same. Hence, (12) can be written as:

$$= \frac{i(K+T)}{K(i+\delta) + U + C + T(i+\delta) + \omega} \tag{13}$$

The save in storage space for the TTHR over the TTMR would be dependable on the values of K , T , and i , if the term $i(K+T)$ is very large compared to ω , then there will be an improvement in the TTHR over the TTMR, otherwise the TTMR has more saving in memory storage over the TTSR specially for a very large value of δ . In this paper, however, The TTHR model is proposed to overcome the complexity of the TTMR model in terms of implementation and query processing. For example, if the number of time-varying attributes i is very large which is normal in temporal database application specially the clinical database applications where very large number of attributes that can vary within a time. Suppose $i = 50$, then a need to create 50 tables, 50 index for the primary key, 50 integrity constraints, and etc. This for one object, if the database schema has more than one temporal object, then the same thing will be repeated, which would result in difficulty to manage the database schema. In addition to that, combing the temporal data of a specific object requires a distinction of relationship recognized as time-varying junction relationship, which is usually costly to put into practice. A parallel research for computing the save in a memory stocking room for the TTMR over the TTSR has been done in [4]. Based on the mathematical equations, several experiments have been carried out; the results explained in more details in our previous work in [18, 25-28]. Several experiments have been carried out at different cost of A_c and freezing $f(A_c)$ at different values, yielded the results that show as saving in memory space range from 80%-90%.

V. Conclusion

TTHR data model has accomplished a saving in memory storage as well as the easy implementation and retrieval of time-varying data in RDBMS. These issues are important in stocking memory space and that is approximately equivalent to or larger than the TTMR. The introduced time-varying information representation is proposed for its straightforwardness, because less information record items will be required to grasp the time-varying features of the temporal information comparing to the TTMR. Furthermore, implementing the TTHR to an available information record application does not necessitate numerous updates in comparison to the TTMR. Other than that, simply have to construct the supplementary association to size the chronological updates of temporal characteristics, but without changing the scheme itself. This contrasts to the TTMR as it needs to decompose the relations and redefine the integrity constraints. A case study for modeling biomedical data (diabetes) has been shown and can be implemented in any RDBMS with a saving in memory usage range from 68-81% over other temporal representations.

Acknowledgment

This project was funded by the Deanship of Scientific Research (DSR) under grant no. G- 27-005-1430. The authors, therefore, acknowledge with thanks to DSR’s technical and financial support.

References

[1] Findler, N. V., & Chen, D. (1973). On the problems of time retrieval of temporal relations causality, and coexistence. International Journal of Computer & Information Sciences, 2, 3, 161-185.

[2] Snodgrass R (1999) Temporal Data Management, IEEE Transactions on Knowledge and Data Engineering, Vol. 11, No. 1, January/February 1999, pp. 36–44).

[3] Date, C. D., Darwen, H., & Lorentzos, N. A. (2003). Temporal data and the relational data model. San Francisco: Morgan Kaufmann.

[4] Novikov, B. A., & Gorshkova, E. A. (2008). Temporal databases: From theory to applications. Programming and Computer Software, 34, 1, 1-6. Pleiades Publishing, Ltd., 2008. Original Russian Text

[5] Tansel, A. U. (2004). On handling time-varying data in the relational data model. Information and Software Technology, 46, 2, 119-126.

[6] Elmasri, R., and Navathe (2000). Fundamentals of Database Systems. 3rd edition. Addison Wesley.

[7] Jensen, C. S., Clifford, J., Gadia, S. K., Segev, A., & Snodgrass, R. T. (1992). A glossary of temporal database concepts. ACM Sigmod Record, 21, 3, 35-43.

[8] Snodgrass, R. T., (2000). Developing Time-Oriented Database Applications in SQL, 1st edition, Morgan Kaufmann Publishers, Inc., San Francisco.

[9] Jensen, C. S., Snodgrass, R. T., & Soo, M. D. (1995). The tsql2 data model (pp. 157-240). Springer US. http://people.cs.aau.dk/~csj/ Thesis/pdf/chapter12.pdf.

[10] Patel, J. (2003). Temporal Database System Individual Project. Department of Computing, Imperial College, University of London, Individual Project, 18-June-2003, http://www.doc.ic.ac.uk/~pjm/ teaching/student_projects/ jaymin_patel.pdf.

[11] Zimányi, E. (2006). Temporal aggregates and temporal universal quantification in standard SQL. ACM SIGMOD Record, 35, 2, 16-21.

[12] Wang, F., Zhou, X., & Zaniolo, C. (2006, April). Using XML to build efficient transaction-time temporal database systems on relational databases. In Proceedings of the 22nd International Conference on Data Engineering, 2006. ICDE'06 (pp. 131-131). IEEE.

[13] A-Qustaishat, M. (2001). A visual temporal object-oriented model embodied as an expert C++ Library. ADVANCES IN MODELLING AND ANALYSIS-D-,6, 3/4, 3-43.

[14] Bohlen, M. H., Busatto, R., & Jensen, C. S. (1998, February). Point-versus interval-based temporal data models. In Proceedings of 14th International Conference on Data Engineering, (pp. 192-200). IEEE.

[15] Dyreson, C., Grandi, F., Käfer, W., Kline, N., Lorentzos, N., Mitsopoulos, Y., ... & Wiederhold, G. (1994). A consensus glossary of temporal database concepts.ACM Sigmod Record, 23, 1, 52-64.

[16] Tansel, A. U. (2006). Modeling and Querying Temporal Data. Idea Group Inc.

[17] Tansel, A. U. (2004). Temporal data modeling and integrity constraints in relational databases. In Computer and Information Sciences-ISCIS 2004 (pp. 459-469). Springer Berlin Heidelberg.

[18] Halawani, S. M., & Romema, N. A. (2010). Memory storage issues of temporal database applications on relational database management systems. Journal of Computer Science, 6, 3, 296.

[19] Atay, C. (2016). An attribute or tuple timestamping in bitemporal relational databases. Turkish Journal of Electrical Engineering & Computer Sciences. (2016) 24: (pp. 4305 – 4321). doi:10.3906/elk-1403-39.

[20] Noh, S.Y., Gadia, S.K. and Jang, H., (2013). Comparisons of three data storage models in parametric temporal databases. Journal of Central South University, 20(7), pp.1919-1927.

[21] Kvet, M., Matiako, K. and Kvet, M., (2014). Transaction management in fully temporal system. In Computer Modelling and Simulation (UKSim), 2014 UKSim-AMSS 16th International Conference on (pp. 148-153). IEEE.

[22] Snodgrass R, Ahn I. Performance evaluation of a temporal database management system. Commun ACM 1986; 15:96-107.

[23] Arora, S. (2015). A comparative study on temporal database models: A survey. In Advanced Computing and Communication (ISACC), 2015 International Symposium on (pp. 161-167). IEEE.

[24] Anselma, L., Stantic, B., Terenziani, P., and Sattar, A. (2013). Querying now-relative data. Journal of Intelligent Information Systems, 41(2), 285-311.

[25] Halawani, S.M., AlBidewi, I., Ahmad, A.R. and Al-Romema, N.A., 2012. Retrieval optimization technique for tuple timestamp historical relation temporal data model. Journal of Computer Science, 8(2), p.243.

[26] Nashwan Alromema, Mohd Shafry Mohd Rahim and Ibrahim Albidewi, “A Mathematical Model for Comparing Memory Storage of Three Interval-Based Parametric Temporal Database Models” International Journal of Advanced Computer Science and Applications(ijacsa), 8(7), 2017. http://dx.doi.org/10.14569/IJACSA.2017.080741.

[27] Alromema, N.A., Rahim, M.S.M. and Albidewi, I., 2016. Temporal Database Models Validation and Verification using Mapping Methodology. VFAST Transactions on Software Engineering, 11(2), pp.15-26.

[28] Ab Rahman Ahmad, N.A., Rahim, M.S.M. and Albidewi, I., 2015. Temporal Database: An Approach for Modeling and Implementation in Relational Data Model. Life Science Journal, 12(3).

[29] Jones S, Mason P, Stamper R (1979) A Relational Specification Language for Complex Rules. Information Systems, 4(4):293{305, November 1979}.

[30] Navathe S, Ahmed R (1989) A Temporal Relational Model and a Query Language. Information Sciences, 49:147{175, 1989}.

[31] Ariav G (1986) A Temporally Oriented Data Model. ACM Transactions on Database Systems, 11(4):499{527, December 1986}.

[32] Sarda N (1990) Extensions to SQL for Historical Databases. IEEE Transactions on Knowledge and Data Engineering, 2(2):220{230, June 1990}.

[33] Snodgrass R (1987) The Temporal Query Language TQUEL. ACM Transactions on Database Systems, 12(2):247{298, June 1987}.

[34] Noh S, Gadia S (2008) Benchmarking temporal database prototypes with interval-based and temporal element-based timestamping. Journal of Systems and Software, 81(11):1931–1943.

[35] Gadia S, (1992) A Seamless Generic Extension of SQL for Querying Temporal Data. Technical Report TR-92-02, Computer Science Department, Iowa State University, May 1992.

[36] https://www.sophe.org/wp-content/uploads/2017/01/Diabetes_toolkitrevised.pdf.

[37] https://www.joslin.org/research/diabetes-research-center.

Artificial Neural Network based Emotion Classification and Recognition from Speech

Mudasser Iqbal[1]*, Syed Ali Raza[2]
Department of Computer Science
Institute of Southern Punjab (ISP) Multan, Pakistan[1]
Government College University (GCU)
Lahore, Pakistan[2]

Furqan Majeed[4]
Department of Computer Science
Institute of Southern Punjab (ISP)
Multan, Pakistan

Muhammad Abid[3]
Department of Computer Science
Govt. Postgraduate College
Layyah, Pakistan

Ans Ali Hussain[5]
Department of Computer Science
University of Agriculture
Faisalabad, Pakistan

Abstract—**Emotion recognition from speech signals is still a challenging task. Hence, proposing an efficient and accurate technique for speech-based emotion recognition is also an important task. This study is focused on four basic human emotions (sad, angry, happy, and normal) recognition using an artificial neural network that can be detected through vocal expressions resulting in more efficient and productive machine behaviors. An effective model based on a Bayesian regularized artificial neural network (BRANN) is proposed in this study for speech-based emotion recognition. The experiments are conducted on a well-known Berlin database having 1470 speech samples carrying basic emotions with 500 samples of angry emotions, 300 samples of happy emotions, 350 samples of a neutral state, and 320 samples of sad emotions. The four features Frequency, Pitch, Amplitude, and formant of speech is used to recognize four basic emotions from speech. The performance of the proposed methodology is compared with the performance of state-of-the-art methodologies used for emotion recognition from speech. The proposed methodology achieved 95% accuracy of emotion recognition which is highest as compared to other states of the art techniques in the relevant domain.**

Keywords—Emotion States; ANN; BR; BRANN; emotion classifier; speech emotion recognition

I. Introduction

In the modern age of technology, Emotion recognition from speech is a hot research topic in the field of speech signal processing [1]. The objective of emotion recognition from the speech is to make Human-Computer Interaction (HCI) more natural and human friendly [2] [3]. However, there is a gap between humans and computers, that computers act logically and humans act logically as well as emotionally. This gap makes the computers less compatible with humans. To reduce this gap, to make the interface easier to use, and to develop more understanding between humans and machines, it is necessary for the machines to understand and act according to human emotions.

There are some well-known emotions like anger, happy, sad, and neutral [4] that can affect speech signals. There are different features of sound like frequency, amplitude, pitch, and format that has been in use for emotion recognition from speech signal [5]. Researchers have been used different approaches such as wavelet-based feature, Mel frequency cepstral coefficient (MFCC), and linear prediction cepstral coefficient (LPCC) [6]. However, Mel-Frequency cepstral coefficients (MFCC) is the most used feature for emotion recognition from speech [7]. The objective of this research is to review emotion recognition techniques with recognition accuracy and to produce an emotion recognition system to recognize four basic emotions anger, sad, happy, and neutral using speech signals.

Emotions recognition can be elaborated as the extraction of emotion from speech signals to make human-computer interaction (HCI) more efficient and convenient [8]. Various techniques are in use for emotion recognition that includes feature selection and extraction and then applying classifier. Hidden Markov model (HMM), Gaussian mixture model (GMM), Support vector machine (SVM), and Artificial neural network (ANN) are the classifiers that can be used for emotion recognition [9], however, the success rate of emotion recognition depends upon features and training algorithm. Training and testing are the two phases in the supervised machine learning approach. During the training phase, the neural network is created and trained by providing a large number of training data and training algorithms with a random number of neurons [10]. During the testing phase, the test data set is provided and its features are extracted and matched with the trained model [10]. Accurate speech recognition depends upon the training algorithm and machine learning accuracy.

II. Related Work

Spectral feature-based emotions classification using Gaussian Mixture Models (GMMs) a technique of acoustic-phonetic approach was proposed by [11]. A private database for emotional speech is used where specific content was recorded by a male and a female actor voices with four basic emotions anger, happy, sad, and neutral in acted speech corpus. For the real speech corpus, samples of single male voices were taken. The four emotions: sad, happy, anger, and neutral were selected for the classification of real speech corpus. 75% of samples were used for training and the rest 25% were used for testing. Gaussian Mixture Model (GMM) was used for the emotions classification because it was considered better for performing clustering. It produced 85% average correct classification for the real speech samples in case of separate gender voices. However, it was observed that when the voices of two males or two females or male and female were mixed, the performance of the purposed system decreased which was the sign that the system was speaker-dependent as well as gender-dependent.

Emotion recognition from the human speech is still a complex task because of the ambiguity in classifying the natural and acted emotion in a human speech explained by [12]. They [12] applied several machine learning techniques including K-nearest neighbor (KNN) and artificial neural network (ANN) to create recognition agents. The agent was able to recognize five emotional states including normal, happiness, anger, sadness, and fear with the accuracy of 55-75%, 60-70%, 70-80%, 75-85%, and 35-55%, respectively. The author in [12] proposed a modified MFCC approach by separating male and female data into separate frames to increase the accuracy of emotion recognition. By applying this modified MFCC approach, before breaking down the speech sample into frames, classification of the speech sample of male and female was done and then compared with the database. So, the overall success rate of the standard approach was 54.54% whereas the modified MFCC approach produced a 63.63% overall successive rate.

Assigning human-like properties e.g. watching, interpreting, and producing effective features is referred to as affective computing which enables the machine to recognize human emotion and to react accordingly. To obtain this goal, a study is conducted by [13]. They have divided the emotion recognition system into feature extraction and feature classification phases. An acted database which is the Berlin Database of emotional speech of 5 male and 5 female actors is used. Three categories of the audio file which were sad, happiness, and neutral states of person are used to reveal this study. At the feature selection phase, a combination of energy, skewness, and MFFC is used. They have used a total of 354 instances with 121 instances of happiness, 117 instances of sadness, and the rest 116 instances of neutral states to conduct this study. A variant number of neurons at the hidden layer of vanilla Artificial Neural Network is used to classify the selected features. The experiment results showed 72% accuracy with 45 neurons at the hidden layer.

Many classification methods have been applied for emotion recognition from speech such as Support Vector Machine (SVM), Hidden Markov Model (HMM), K-Nearest Neighbors (KNN), and Artificial Neural Network (ANN) [14].

A classification technique (optimized Support Vector Machine) for emotion recognition from the speech was proposed by [15]. In that work, they used four emotions: anger, happiness, sadness, and surprise. For the experimental purpose, the voice of a group of non-professionals was recorded having four basic emotions. The experiment was speaker and gender independent, therefore, 5 males and 9 female speakers with age group about 20 took part in recording the emotional speeches. Fast Fourier Transform (FFT) was used to transform every frame of N samples into a frequency spectrum. After that, an optimized support vector machine (SVM) was established to perform two-class pattern classification. Four kernel functions which were Linear, Polynomial, Radial basis function (RBF), and Sigmoid function to evaluate the performance of optimized support vector machine were used for emotion recognition from speech. According to the [15] research, the Radial basis function (RBF) produced 71.89% accuracy at the training set and 88.75% accuracy at the testing set.

A study on emotion recognition from speech was directed by [16] to design and propose an efficient classifier for emotion recognition from speech. Their work for the recognition of emotions from the speech was based on a discrete emotion classification system. They constructed a model for emotion recognition from speech based on support vector machine (SVM) and artificial neural network (ANN) respectively. The effect of emotion recognition of feature reduction was compared respectively for both support vector machine (SVM) and artificial neural network (ANN). According to the [16], experimental results showed that artificial neural networks (ANN) produced 75 % testing accuracy and support vector machine (SVM) produced 85% accuracy. It was proposed that the support vector machine (SVM) is slightly better than the artificial neural network (ANN).

III. Proposed Methodology

The importance of emotion recognition from an audio signal is increasing to make human-machine interaction more efficient. It is based on depth analysis of speech signal, feature extraction that contains emotional information, and taking appropriate pattern recognition techniques to recognize the emotional state. The first step is to extract features like pitch, frequency, timbre, and amplitude from the audio signal to train a neural network. The second step is to give an audio input to a machine to compare with the store information obtained at the training time [17]. Speech emotion recognition system is shown in below-given Fig. 1 which includes a sensor module to receive speech data as input, a pre-processing module to remove noise from the data, and a feature extraction module to extract features (frequency, formant, pitch, and amplitude) from the processed data and a training module (Neural Network) with training function.

A. Preprocessing of Speech Signal

Before converting the auditory signals into numerical values, there exist some unnecessary data that get mixed into the original data and affect the original values. The noise of the environment is one of the major factors that get the mix into the actual voice signals and changes their characteristics [18]. Therefore, it is necessary to remove the noise of the given data. This feature extraction module is further subdivided into two modules are Noise removal system and Feature extraction. The job of this module is to remove the noise from the actual data. Because removal of noise is important to get the correct results [11].

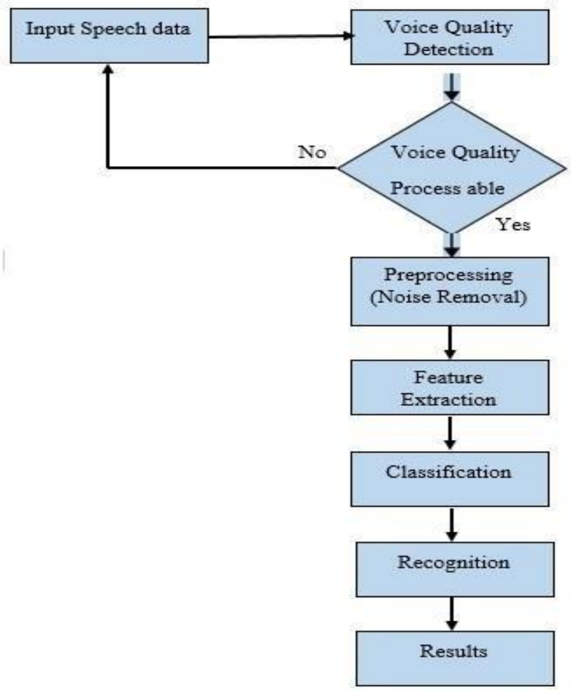


Fig. 1. Flow of Proposed Emotion Recognition System.

B. Feature Extraction

After removing the noise, the task is to extract the features from the data. A sound signal has a lot of features but after the literature review, it came to know that the four features (frequency, amplitude, pitch, and formant) can play a vital role in emotion detection. These four features are considered for emotion detection. The characteristics of these features are given below.

1) FREQUENCY: Frequency, additionally called wave recurrence, is an estimation of the absolute number of vibrations or motions made inside a specific unit of time [19]. As the number of waves per unit time increases the frequency increases which can be written as:

$$Ƒ = 1/T \tag{1}$$

The Time interval in which a wave completes its one cycle is called the time period which is denoted by T and the number of cycles in one second is called frequency which is the reciprocal of the time period as described in equation (1).

2) FORMANT: A formant plays a very vital role in emotion recognition from speech and is a widely used feature for emotion recognition [19]. There are many formants and each formant has a different frequency, roughly assumed one in each 1000Hz band. In the vocal set, every formant corresponds to a specific character. Formants are displayed as dark bands and can be seen very clearly in a wideband spectrogram. As much as the formant is darker, it can be reproduced in the spectrogram and has more energy (stronger it is or the more audible it is). The equation for formant is:

$$L = C/4F \tag{2}$$

Where "C" is the speech of sound (340.29 m/s) and "F" is the first formant of the frequency.

3) PITCH: Pitch is also a widely used feature in emotion recognition and it is a perceptual property of sounds that permits their ordering on a Frequency related scale or pitch is the nature of frequency that makes it conceivable to acknowledge sounds as" higher" and" lower". Pitch is considered as the hear-able quality of sound recurrence as per which sounds can be ordered on a scale from low to high [20] [21]. The term" high" pitch is considered as exceptionally quick wavering, and the term" low" pitch portrays the more slow swaying.

$$\text{Pitch} \propto 1/T \tag{3}$$

As described in equation (3), pitch directly depends upon frequency. This shows that the greater is the frequency of a sound, the greater will be its pitch.

4) AMPLITUDE: Amplitudes can be defined either as instantaneous values or mostly as peak values. Amplitude is referred to as the fluctuation or displacement of a wave. With sound waves, the amplitude is the loudness of the sound and the extent to which air particles are displaced [22]. Also analyzed the feature of Amplitude, Pitch, and Formant in their paper.

$$X = X_m \mathrm{Sin}(\omega t+\varphi) \tag{4}$$

In equation (4), X is the instantaneous value of displacement of a wave from the mean position and X_m is amplitude where ($\omega t+\varphi$) is called the phase of the motion.

C. Bayesian Regularization

Bayesian Regularization (BR) typically consumes less memory and more time but it has a good generalization to process difficult and noisy datasets. Training of Bayesian Regularization algorithm automatically stops according to the adaptive weight minimization regularization. Bayesian Regularization algorithm is more accurate as compare to Levenberg Marquardt (LM) and Scale Conjugate Gradient (SCG) [23]. It prevents overtraining and provides an efficient criterion for stopping the training process. This efficiency of the BR algorithm makes it a more adaptive algorithm for training a neural network to perform emotion recognition from speech. We can write Bayesian theorem as:

$$P\left(\frac{X}{Y}\right) = P\left(\frac{X}{Y}\right)\left(\frac{P(X)}{P(Y)}\right) \tag{5}$$

Bayesian regularized artificial neural network (BRANNs) is stronger than standard back-propagation networks and can decrease or expel the requirement for long cross-validation. It changes over a nonlinear regression into a well-structured linear regression as:

$$Y = B0X / (B1 + X) \quad \text{Nonlinear regression} \tag{6}$$

$$Y = a + bX + u \quad \text{Linear regression} \tag{7}$$

Where Y is the variable that is to be predicted and X is used to predict Y as well as a is the intercept, b is the slope and u is the regression residual. Therefore, Bayesian regularized artificial neural network (BRANN) can be expressed in the equation as:

$$D(Z) = \beta \sum_{x=1}^{N_C} \quad [Y_i - f(B_i)]^2 + \alpha \sum_{j=1}^{N_Z} \quad {Z_j}^2 \qquad (8)$$

Where Nz is the number of weights. Given starting values of the hyperparameters α and β the cost function, D (Z), is reduced regarding the weights Z. A re-estimate of α and β is made by amplifying the proof. The initial probability over the weight, Z, can be composed as:

$$P(Z \mid \alpha, H) = \frac{1}{\mathrm{C_Z}} exp(-\alpha E_{\mathrm{Z}}) \qquad (9)$$

$$E_Z = \sum_{x=1}^{N_Z} \quad {Z_j}^2 \; being\ the\ error\ of\ the\ weights. \qquad (10)$$

With

Similarly, the probability of errors can be described as

$$P(E)|Z, \beta, H) = \frac{1}{W_E(\beta)} exp(-\beta E_D) \qquad (11)$$

With

$$E_Z = \sum_{j=1}^{N_Z} \quad [y_i - f(X_j)]^2 \; being\ the\ error\ of\ the\ data \qquad (12)$$

The most common function used for Bayesian estimation is the mean square error (MSE) which can be defined as:

$$\text{MSE} = E[(\hat{\theta}\,(x) - \theta)^2] \qquad (13)$$

Where the expectation is taken over the joint distribution of θ and x.

D. Artificial Neural Network

Artificial Neural Network (ANN) is a scientific model, which works similar to the human neural system [24] [25]. A two-layer feed-forward network is trained with a Bayesian Regularization algorithm. A feed-forward neural network, which consists of multiple numbers of neurons as a processing unit and is a biologically inspired classification algorithm [26]. Each processing unit in a layer is connected with all processing units of the previous layer. Speech-based emotion classification is examined in three main steps: Data preprocessing and noise removal, feature extraction, emotion classification and recognition, and then output.

Artificial Neural Network (ANN) is trained using emotional speech data taken from the Berlin Database of Emotional Speech of the Technical University of Berlin, Institute of Speech and Communication, department of communication science [27] [28]. This Database Consists of 3300 utterances spoken by the different actors in happy, angry, sad, disgust, boredom, fear, and normal ways. As this work aims at 4 emotions States as Angry, Sad, Normal, and Happy, therefore, four utterances are filtered and used the Feature Extraction Module to generate the Dataset to train Bayesian Regularized Artificial Neural Network (BRANN). A total of 1470 samples of Emotions (500 samples of angry emotion, 300 samples for happiness, 350 samples for natural, and 320 samples for sorrow) are taken to train the proposed ANN. Emotion recognition from the speech is done in three main phases. In the First Phase, Speech data is received and noise is removed to get better results and more accurate feature extraction. In the second Phase, features (Frequency, Pitch, Amplitude, and Formant) are extracted. After extracting the features from auditory data, these features are fed to the neural network. In the third Phase, emotional speech is classified and passed to the recognition phase that produced the final output. As mentioned earlier network consists of seventy-eight layers including seventy hidden layers 4 input layers and 4 output layers. The output of the ith layer becomes the input for the (i+1)the layer. Every time biases are added to the input. After every iteration, the input values change for every neuron because new weights are adjusted after every iteration. This neural network contains seventy-eight layers of neurons including 4 input layers, 4 output layers, and seventy hidden fully connected layers as shown in below Fig. 2.

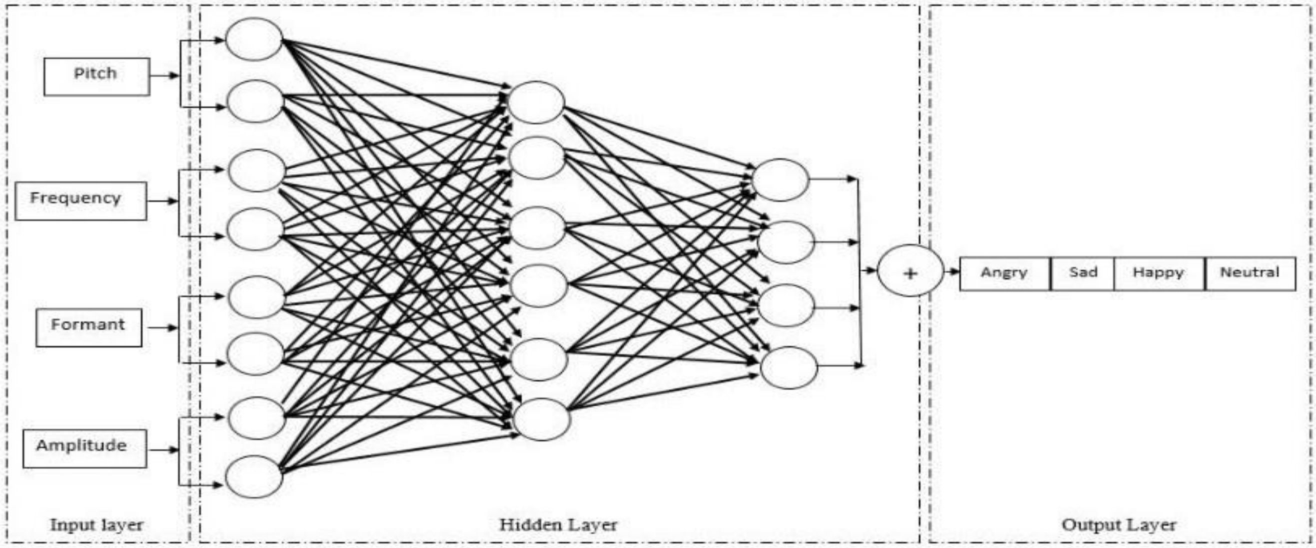


Fig. 2. Artificial Neural Network Framework for Speech-based Emotion Recognition.

IV. Results and Experiments

It has been observed that the emotions of a human change the features of its voice with a specific pattern. For example, when a human gets angry his intensity of voice increases so the frequency increases and in most of the cases, he goes louder to show his anger so the amplitude of his voice signal also increases. Similarly, when someone becomes sad his intensity and amplitude of voice signal also get low. The experiments were executed in Matlab 2017r2, Dell i7 with properties of 8 GB RAM, Windows 10 operating system. The proposed system is trained and tested with 1470 speech datasets taken from the Berlin database carrying four basic emotions. After training the network, testing is done by providing emotional speeches to the network, which are successfully recognized. The following figures show the testing result of the proposed network based on four features (frequency, pitch, formant, and amplitude. The experimental results of speech-based emotion recognition through Bayesian Regularized Artificial Neural Network on the Berlin database of emotional datasets have produced an efficient performance as compared to the state-of-the-art techniques. The proposed methodology is executed in three phases which are briefly explained in Section III.

In the first phase, preprocessing like noise removal is performed on the audio dataset and passed to the next phase. In the second phase, which is also an important phase of the proposed system, features are extracted from the speech data. After the feature (frequency, pitch, amplitude, and formant) extraction the speech data is passed to the third phase, and classification is done using ANN. We have used a 1470 emotional speech dataset to train the Artificial Neural Network with the Bayesian Regularized algorithm due to its less memory and time consumption and then tested with 700 emotional datasets. Results are shown in below-given Fig. 3.

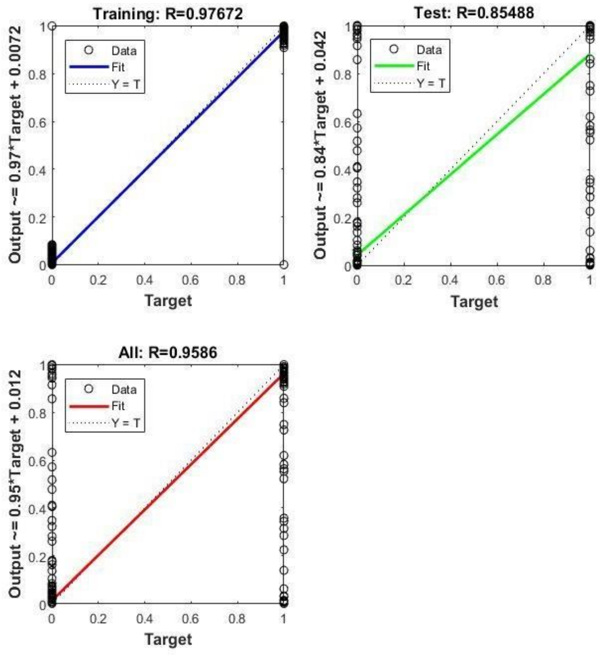


Fig. 3. Regression Plot for Training and Testing of Artificial Neural Network having 70 Neurons at Hidden Layer with BR Algorithm.

As shown in Fig. 3, optimum results are obtained with 70 neurons at the hidden layer. The proposed technique produced 97.6 % training accuracy, 85.4% testing accuracy, and 95.8% overall accuracy of emotion recognition from the speech which is superior when performing a comparison with the state-of-the-art techniques of emotion recognition from speech. Table I and Fig. 4 shows the confusion matrix of the proposed system. The proposed system performance is also compared with the state-of-the-art techniques in "Table II" which shows that the proposed system produced more accurate results than existing techniques.

TABLE I. Performance Comparison of the Proposed System with a Variant Number of Neurons at the Hidden Layer of BRANN

Training Algorithm: **Bayesian Regularization**				
Number of Neurons	**Training Rate**	**Testing Rate**	**Overall Accuracy**	**Best Performance (Error Rate %)**
10	88%	84%	87%	0.04
20	93%	82%	91%	0.02
30	95%	82%	93%	0.01
40	96%	80%	94%	0.01
70	97%	85%	95%	0.007
80	97%	72%	93%	0.02

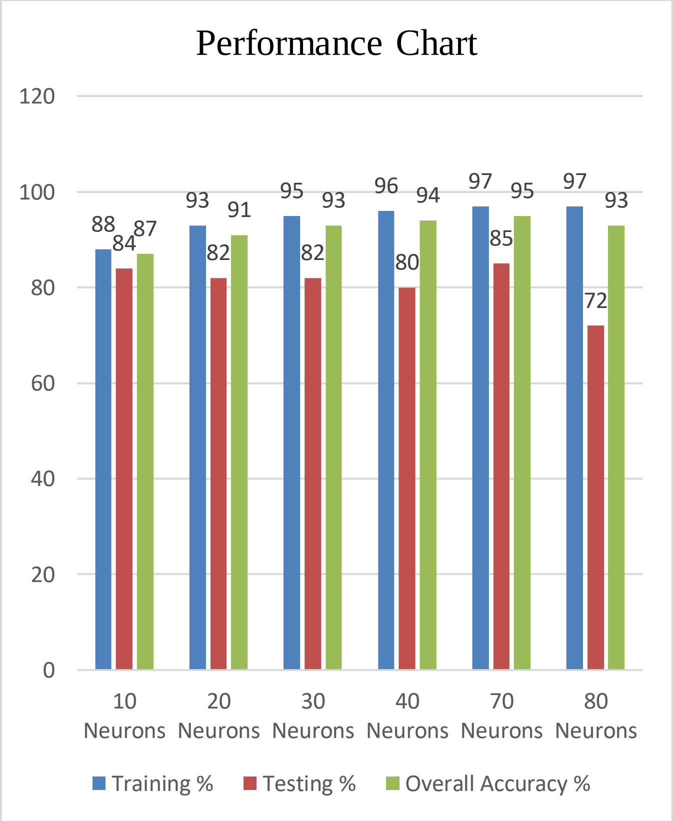


Fig. 4. Performance Comparison of Proposed Technique.

As shown in Table I, and Fig. 4, accuracy in results increased with increasing the number of neurons at the hidden layer but with 80 neurons at the hidden layer, results are significantly decreased therefore the training process is stopped. Optimum results are obtained with 70 number of neurons at the hidden layer. The performance of the proposed system is also compared with state-of-the-art methodologies and it is demonstrated that the performance of the proposed technique is comparatively better and efficient. The following figures show the results of testing the proposed technique by providing an emotional dataset.

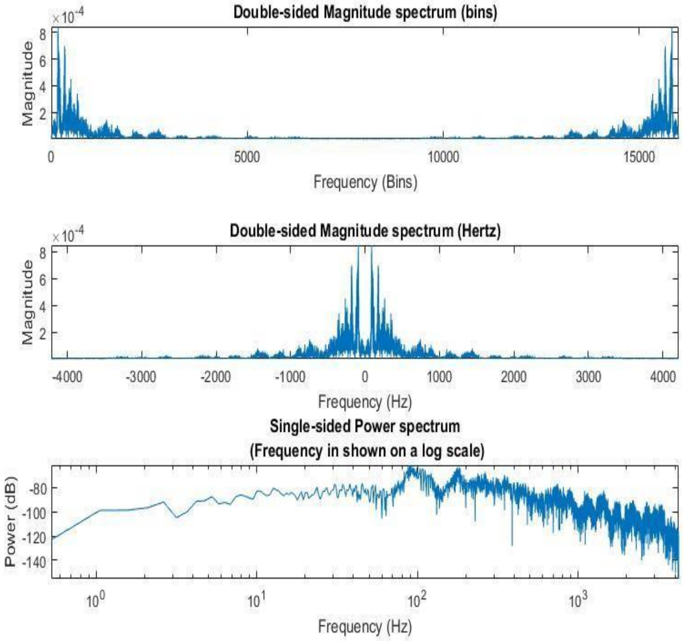


Fig. 5. The Plot of Frequency Feature for Speech Data Containing Angry Emotion.

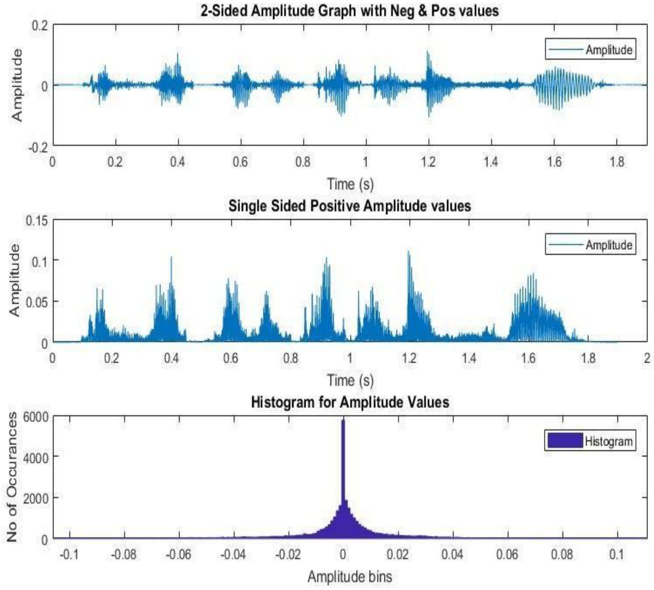


Fig. 6. The Plot of Amplitude Feature for Speech Data Containing Angry Emotion.

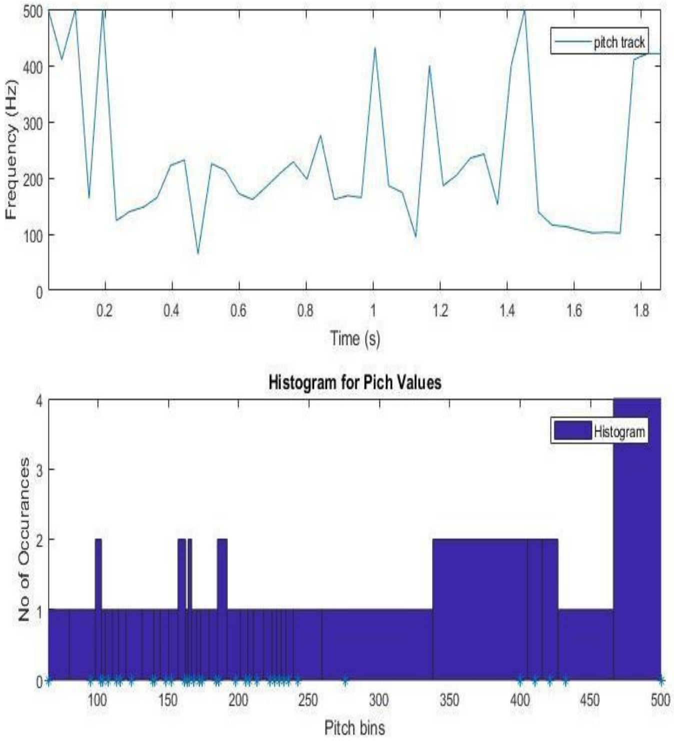


Fig. 7. The Plot of Pitch Feature for Speech Data Containing Angry Emotion.

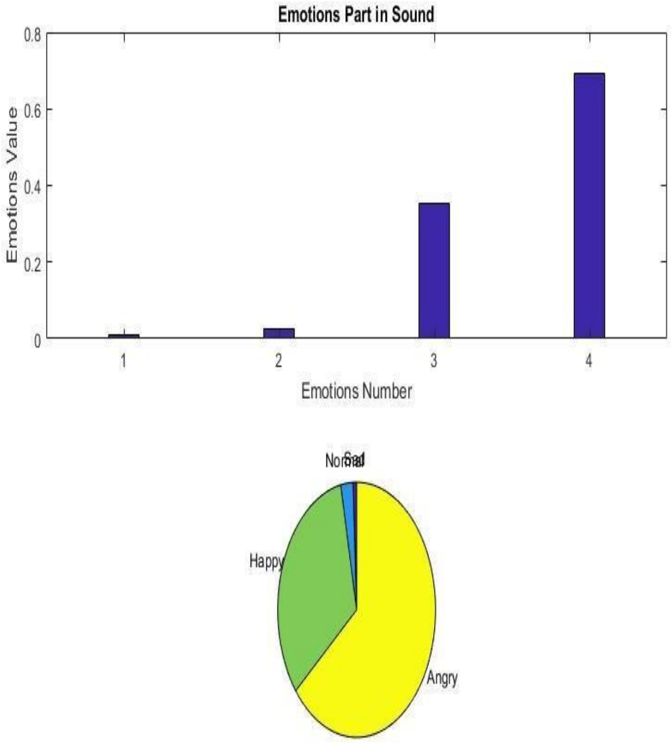


Fig. 8. Angry Emotion Recognized Successfully from Speech Dataset.

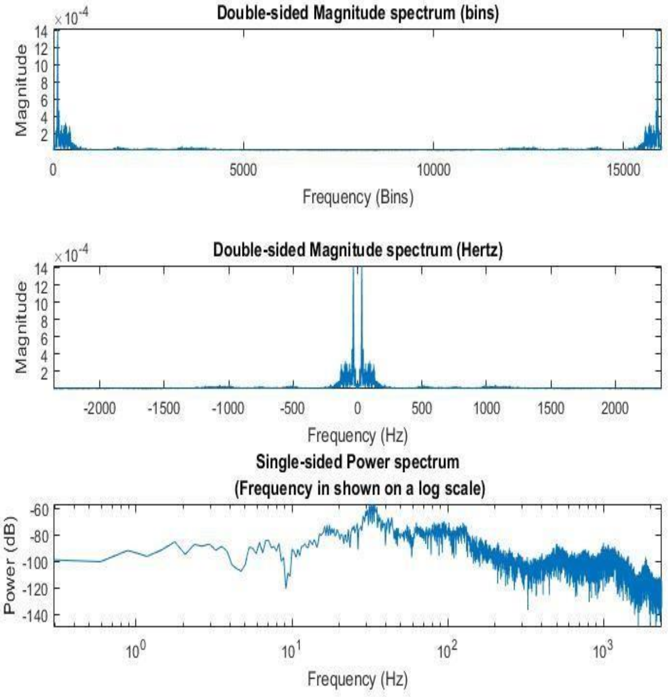


Fig. 9. The Plot of Frequency Feature for Recognizing Happy Emotion.

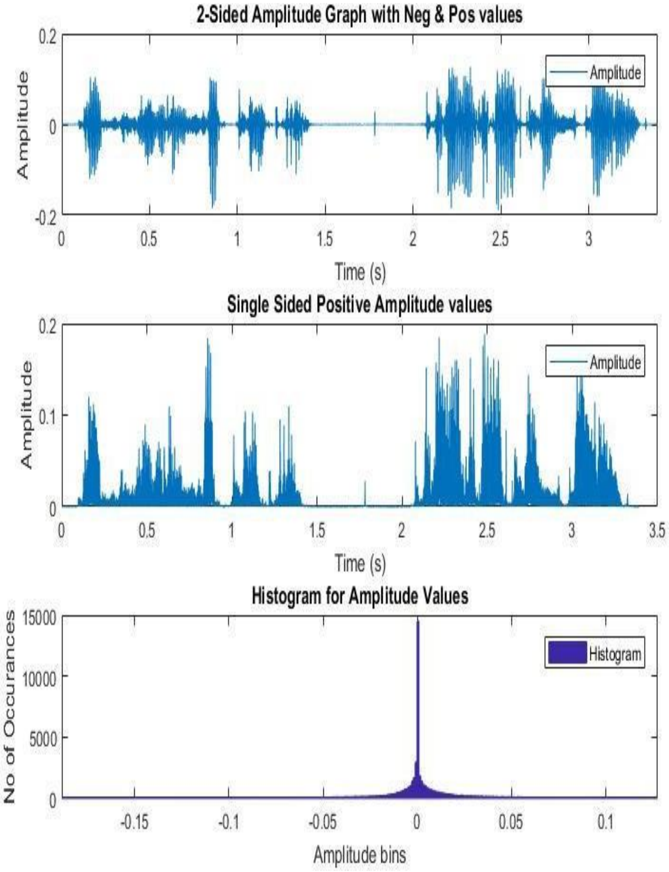


Fig. 10. The Plot of Amplitude Feature for Speech Data Containing Happy Emotion.

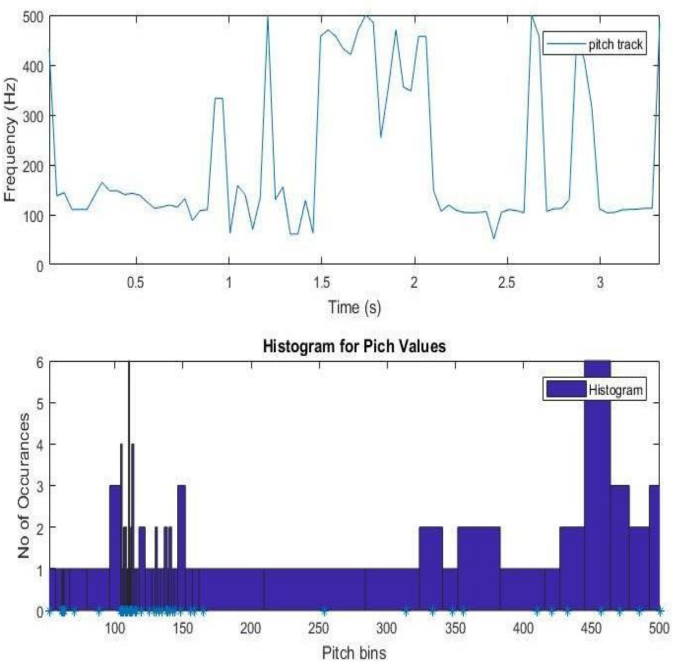


Fig. 11. The Plot of Pitch Feature for Speech Data Containing Happy Emotion.

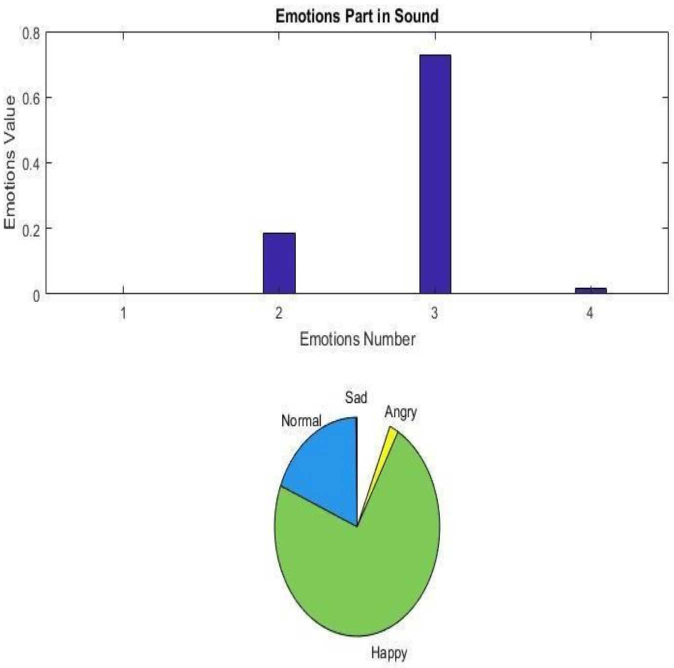


Fig. 12. Happy Emotion Recognized Successfully from Speech Dataset.

As shown in “Fig. 5” to “Fig. 12”, the proposed system has produced accurate results basis on the four features (frequency, pitch, formant, and amplitude). Because formant has only values and can’t be displayed in a plot, only frequency, amplitude, and pitch plots as well as recognition graph is displayed in figures. Sad and Neutral emotions are also recognized successfully which can be shown below given “Fig. 13” to “Fig. 20”.

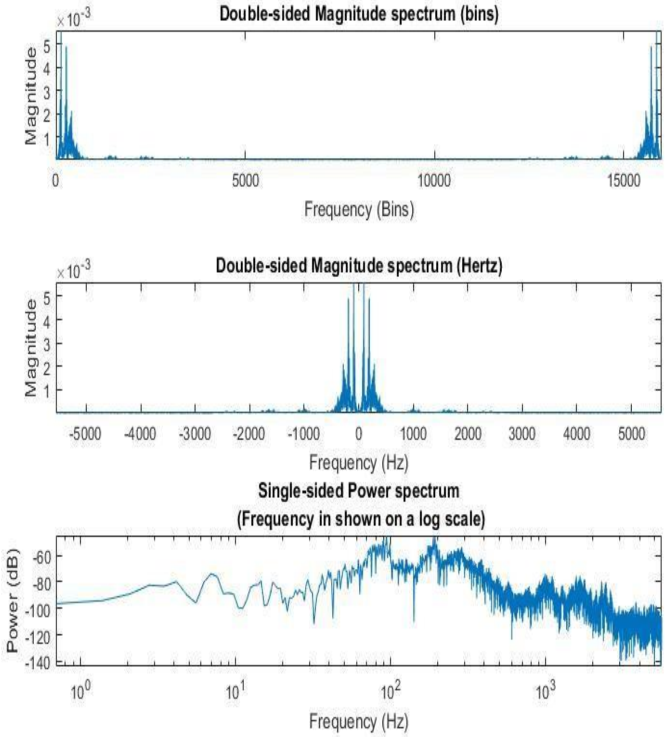


Fig. 13. The Plot of Frequency Feature for Recognizing Neutral Emotion.

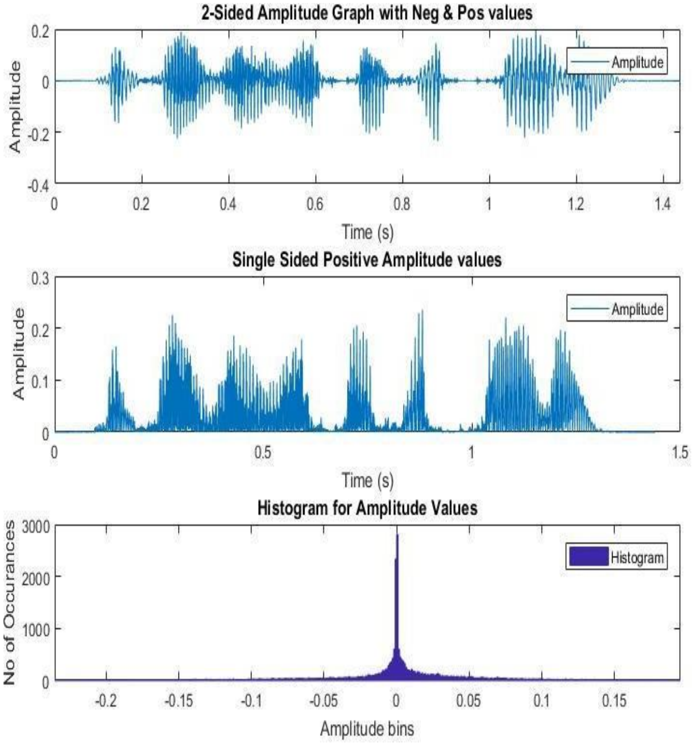


Fig. 14. Amplitude Feature for Speech Data Containing Neutral Emotion.

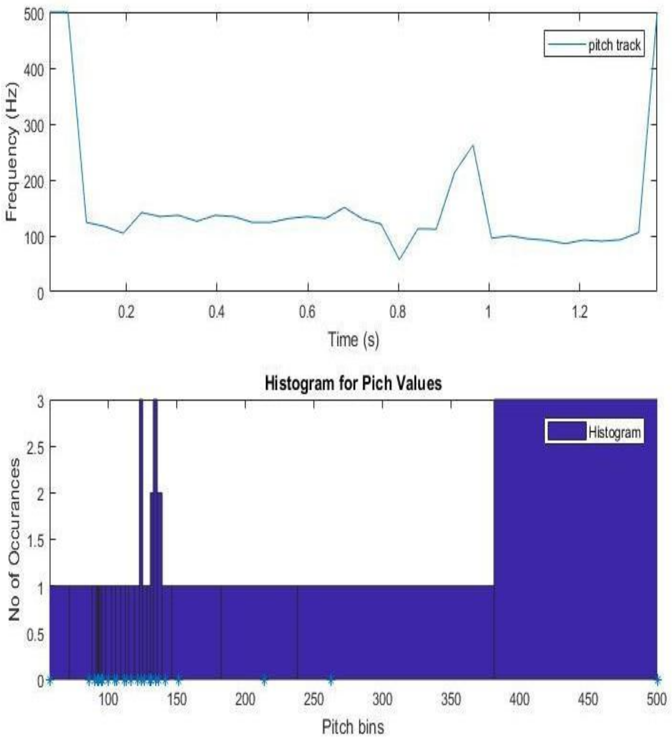


Fig. 15. The Plot of Pitch Feature for Speech Data Containing Happy Emotion.

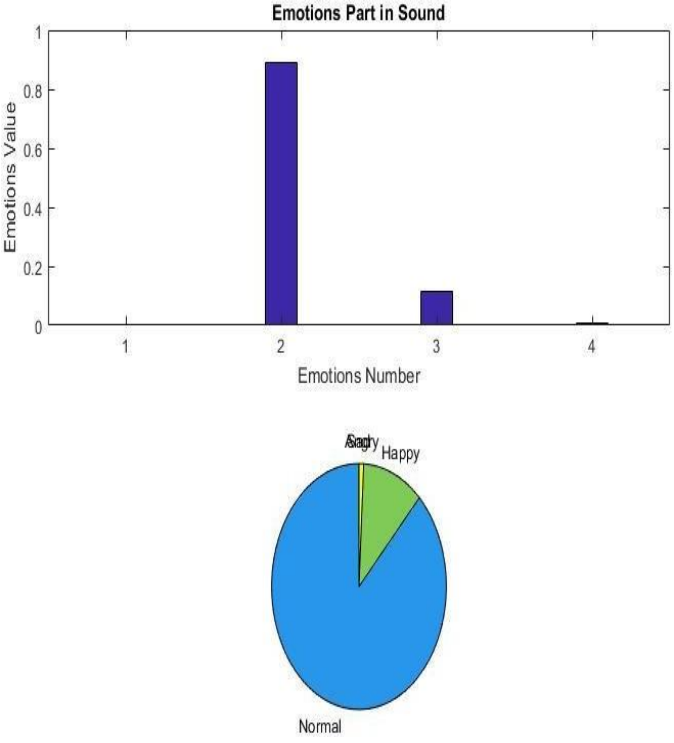


Fig. 16. Neutral Emotion Recognized Successfully from Speech Dataset.

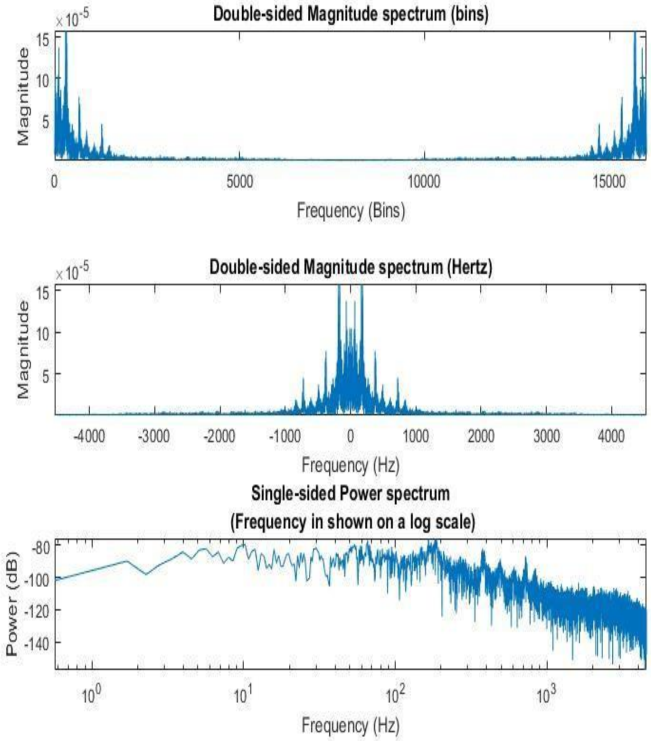


Fig. 17. The Plot of Frequency Feature for Recognizing Sad Emotion.

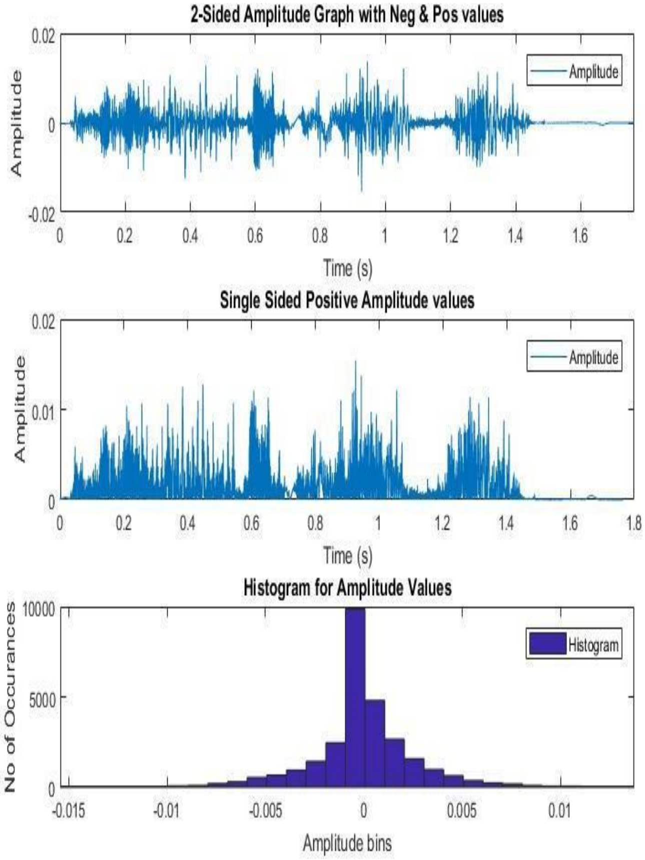


Fig. 18. Amplitude Feature for Speech Data Containing Sad Emotion.

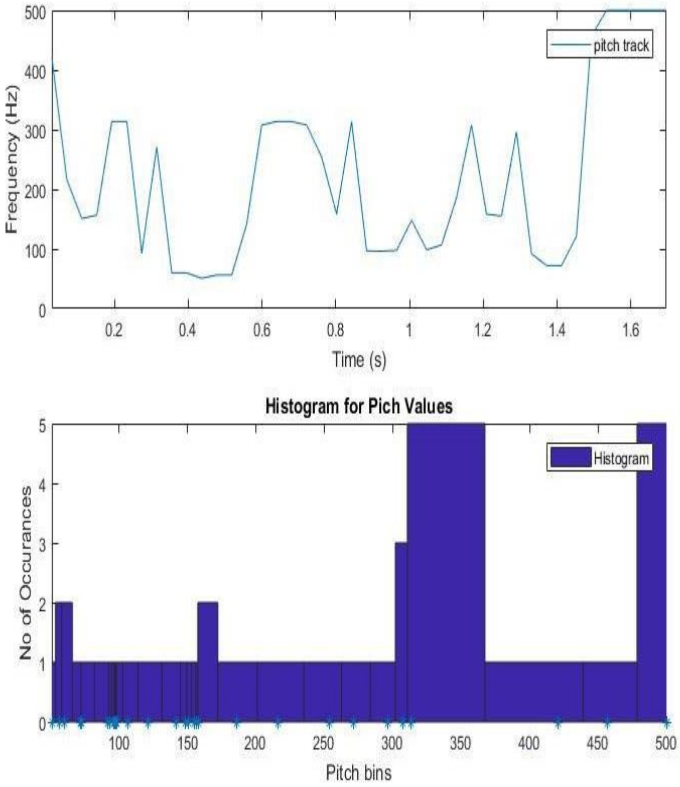


Fig. 19. The Plot of Pitch Feature for Speech Data Containing Sad Emotion.

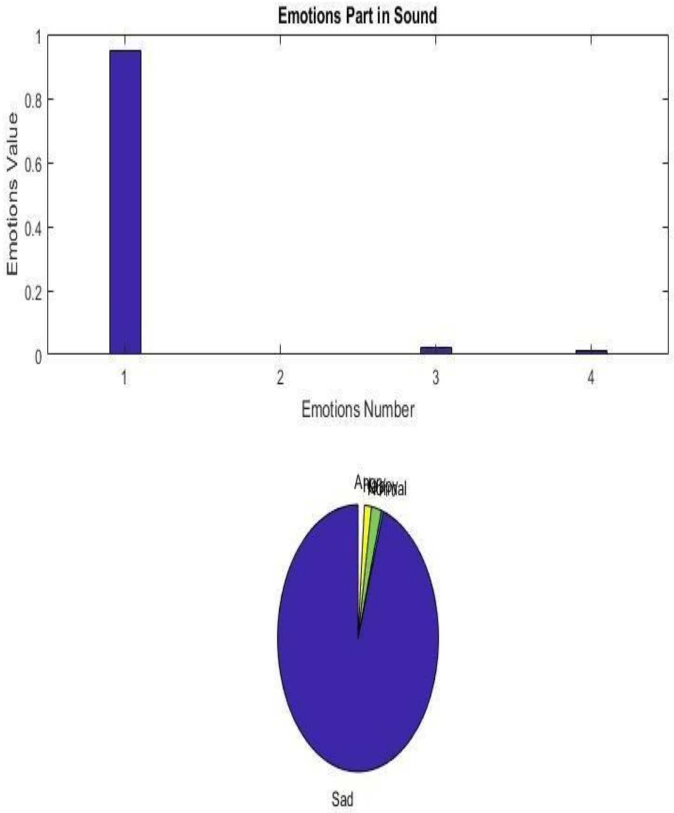


Fig. 20. Sad Emotion Recognized Successfully from Speech Dataset.

TABLE II. COMPARISON OF THE PROPOSED SYSTEM WITH STATE-OF-THE-ART TECHNIQUES

Reference and Year	Methodology	Performance
[11] 2012	Gaussian Mixture Model (GMM)	85% average correct classification for the real speech samples with separate gender voices
[12] 2013	K-nearest neighbor (KNN), artificial neural network (ANN) with modified MFCC approach	54.54% accuracy whereas the modified MFCC approach produced 63.63%.
[13] 2018	vanilla Artificial Neural Network (ANN) with berlin emotion database	72% accuracy with 45 neurons at the hidden layer.
[15] 2016	optimized support vector machine (SVM) with different kernel functions	Radial basis function (RBF) produced 71.89% training accuracy and 88.75% testing accuracy.
[16] 2018	support vector machine (SVM) and artificial neural network (ANN) respectively	artificial neural network (ANN) produced 75 %testing accuracy and support vector machine (SVM) produced 85% accuracy
Proposed System	ANN with Bayesian Regularization	97% training accuracy,85% testing accuracy, and 95% overallaccuracy

V. CONCLUSION

The proposed architecture is based on Artificial Neural Network with a Bayesian Regularization algorithm. An artificial agent is created and trained with Berlin's emotional database having four emotions: angry, sad, neutral, and happiness. The proposed system is tested with 10 neurons at the hidden layer initially, which are increased step by step. The proposed architecture with 70 neurons at the hidden layer, produced 97% training accuracy, 85% testing accuracy, and 95% overall accuracy for the four basic (Angry, Happy, Neutral, and Sad) emotions. This work contributed almost a 5% gain in training accuracy and a 3% gain in testing accuracy using Bayesian Regularization Artificial Neural Network (BRANN). The proposed architecture results are also compared with the state-of-the-art techniques for speech-based emotion recognition. The comparison results show that the proposed system recognizes four basic emotions from speech more accurately than state-of-the-art techniques. Moreover, higher accuracy can be obtained using the combination of more features.

REFERENCES

[1] P. Song, W. Zheng, S. Ou, X. Zhang, Y. Jin, J. Liu and Y. Yu, "Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization," Speech Communication, vol. 83, no. 2016, pp. 34-41, 2016.

[2] A. Milton, S. S. Roy and S. T. Selvi, "SVM Scheme for Speech Emotion Recognition using MFCC Feature," International Journal of Computer Applications, vol. 69, no. 9, pp. 34-39, 2013.

[3] Y. Wang and L. Guan, "Recognizing Human Emotional State From Audiovisual Signals," IEEE TRANSACTIONS ON MULTIMEDIA, vol. 10, no. 5, pp. 936-945, 2008.

[4] A. Joshi and R. Kaur, "A Study of Speech Emotion Recognition Methods," International Journal of Computer Science and Mobile Computing, vol. 2, no. 4, pp. 28-31, 2013.

[5] K. M. Kudiri, G. K. Verma and B. Gohel, "Relative Amplitude based Features for Emotion Detection from Speech," in IEEE International Conference on Signal and Image Processing, Chennai, 2010.

[6] X. MAO, B. ZHANG and Y. LUO, "SPEECH EMOTION RECOGNITION BASED ON A HYBRID OF HMM/ANN," in Proceedings of the 7th WSEAS International Conference on Applied Informatics and Communications, Athens, 2007.

[7] S. Demircan and H. Kahramanlı, "Feature Extraction from Speech Data for Emotion Recognition," Journal of Advances in Computer Networks, vol. 2, no. 1, pp. 28-30, 2014.

[8] S. N. R. A, R. P. Gadhe, R. R. Deshmukh, V. B. Waghmare and P. P. Shrishrimal, "Automatic emotion recognition from speech signals: A Review," International Journal of Scientific & Engineering Research, vol. 6, no. 4, pp. 636-638, 2015.

[9] K. Wang, N. An, B. N. Li and Y. Zhang, "Speech Emotion Recognition Using Fourier Parameters," IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, vol. 6, no. 1, pp. 69-74, 2015.

[10] S. K.Gaikwad, B. W.Gawali and P. Yannawar, "A Review on Speech Recognition Technique," International Journal of Computer Applications, vol. 10, no. 3, pp. 16-23, 2010.

[11] K. S. Rao, T. P. Kumar, K. Anusha, B. Leela, I. Bhavana and S. V. Gowtham, "Emotion Recognition from Speech," International Journal of Computer Science and Information Technologies, vol. 3, no. 2, pp. 3603-3607, 2012.

[12] A. Sapra, N. Panwar and S. Panwar, "Emotion Recognition from Speech," International Journal of Emerging Technology and Advanced Engineering, vol. 3, no. 2, pp. 341-345, 2013.

[13] A. K. Komal Rajvanshi, "An Efficient Approach for Emotion Detection from Speech Using Neural Networks," International Journal for Research in Applied Science & Engineering Technology, vol. 6, no. 5, pp. 1062-1065, 2018.

[14] R. D.Shah and D. Anil.C.Suthar, "Speech Emotion Recognition Based on SVM Using MATLAB," International Journal of Innovative Research in Computer and Communication Engineering, vol. 4, no. 3, pp. 2916-2920, 2016.

[15] B. Yu, H. Li and C. Fang, "Speech Emotion Recognition based on Optimized Support Vector Machine," JOURNAL OF SOFTWARE, vol. 7, no. 12, pp. 2726-2732, 2012.

[16] X. Ke, Y. Zhu, L. Wen and W. Zhang, "Speech Emotion Recognition Based on SVM and ANN," International Journal of Machine Learning and Computing, vol. 8, no. 3, pp. 198-201, 2018.

[17] A. Joshi and R. Kaur, "A Study of Speech Emotion Recognition Methods," International Journal of Computer Science and Mobile Computing, vol. 2, no. 4, p. 28–31, 2013.

[18] N. J. Gogoi and J. Kalita, "Emotion Recognition from Acted Assamese Speech," International Journal of Innovative Research in Science, Engineering and Technology, vol. 4, no. 6, pp. 4116-4121, 2015.

[19] S. Bhadra, U. Sharma and A. Choudhury, "Study on Feature Extraction of Speech Emotion Recognition," ADBU-Journal of Engineering Technology, vol. 4, no. 1, pp. 7-9, 2016.

[20] Y. Pan, P. Shen and L. Shen, "Speech Emotion Recognition Using Support Vector Machine," International Journal of Smart Home, vol. 6, no. 2, pp. 101-106, 2012.

[21] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," Speech Communication, vol. 48, no. 2006, p. 1162–1181, 2006.

[22] X. m. Cheng, P. y. Cheng and L. Zhao, "A Study on Emotional Feature Analysis and Recognition in Speech Signal," in IEEE 2009 International Conference on Measuring Technology and Mechatronics Automation, Zhangjiajie, 2009.

[23] A. Payal, C. Rai and B. Reddy, "Comparative analysis of Bayesian regularization and Levenberg-Marquardt training algorithm for

localization in wireless sensor network," in IEEE 15th International Conference on Advanced Communications Technology (ICACT), PyeongChang, 2013.

[24] A. Gupta and A. Joshi, "Speech Recognition using Artificial Neural Network," in IEEE 2018 International Conference on Communication and Signal Processing (ICCSP), Chennai, 2018.

[25] I. A. Maaly and M. El-Obaid, "Speech Recognition using Artificial Neural Networks," in IEEE 2006 2nd International Conference on Information & Communication Technologies, Damascus, 2006.

[26] K. S and C. E, "A Review on Automatic Speech Recognition Architecture and Approaches," International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 9, no. 4, pp. 393-404, 2016.

[27] A. Milton, S. S. Roy and S. T. Selvi, "SVM Scheme for Speech Emotion Recognition using MFCC Feature," International Journal of Computer Applications, vol. 69, no. 9, pp. 34-38, 2013.

[28] "Berlin Emotional Speech Database," [Online]. Available: http://www.emodb.bilderbar.info/download/.

Problematic Use of Mobile Phones during the COVID-19 Pandemic in Peruvian University Students, 2020

Rosa Perez-Siguas[1], Randall Seminario-Unzueta[2]
Hernan Matta-Solis[3], Melissa Yauri-Machaca[4], Eduardo Matta-Solis[5]
Research Directorate, Universidad María Auxiliadora, Lima, Perú[1, 3]
Research Unit, Universidad María Auxiliadora, Lima, Perú[2]
Research and Technology Direction, Business on Making Technologies, Lima, Perú[4]
Research Unit of the Faculty of Health Sciences, Universidad María Auxiliadora, Lima, Perú[5]

Abstract—**The problematic use of mobile phones during the COVID-19 pandemic has been predicted as a mental alteration in university students due to confinement due to the health crisis that exists in our country and in the world, therefore, the objective of the study is to determine the problematic use of mobile phones during the COVID-19 pandemic. COVID-19 in Peruvian university students, 2020. This is a quantitative, non-experimental, descriptive, and cross-sectional study, with a population of 163 Peruvian university students, who answered a questionnaire of sociodemographic data and the Mobile Phone Problem Use Scale. In the results where it can observe regarding the problematic use of mobile phones that 103 (63.2%) of the university students have a high problematic use, 59 (36.2%) medium problematic use and 1 (0.6%) use low problematic use. In conclusion, programs on mental health should be carried out during the COVID-19 pandemic in university students.**

Keywords—Mental health; pandemic; coronavirus; mobile phones

I. Introduction

Worldwide, mobile phones have become an important electronic device of the population as a means of communication [1], although excessive use has generated controversy and addictions within the young population, where they use mobile phones just for socialize by social networks and entertainment [2][3].

Although it is true, we are in a global health crisis due to the coronavirus pandemic (COVID - 19), today mobile phones are being used in a very excessive way, generating difficulties in their physical and mental health in the population [4], seeking factors such as anxiety, insomnia and stress, hindering the quality of life that had previously before the COVID -19 pandemic [5].

Likewise, the use of mobile phones has allowed the population to find a way to be entertained by the COVID-19 pandemic, but it also produces changes in the person's mood such as anger, sadness, depression, impulsivity [6], which are manifested as being isolated and quarantined as a result of the pandemic [7]. In such a way that it compromises the state of health of people, although it primarily affects the youngest because they are the ones who use mobile phones the most [8][9].

If the COVID-19 pandemic has caused young people to use their phones more, generating emotional and intellectual changes that compromise their academic performance affecting their daily lives, consequently young people will present long or short mental changes once the COVID-19 pandemic has subsided [10].

In China [11], a study was carried out in 754 adolescent students where 48.10% were women who had experience in the use of the mobile phone, where it was found that the time that adolescents spend on the mobile phone was related more to mobile addictions and psychological anguish, therefore, Chinese adolescents tended to have addiction in relation to the use of mobile phones for using them for a long time daily.

Similarly in China [12], a study was conducted in 847 university students, 51.2% men and 48.8% women, on the relationship between excessive use of mobile phones and alexithymia, where they observed that alexithymia predicts addiction to excessive use of mobile phones, whether directly or indirectly for entertainment.

In Spain [13], a study was carried out in 124 university students of the nursing career 79% were women and 21% were men, in relation to the use of mobile phones and nomophobia, observing that the male gender showed a lower tendency than the women in the use of mobile phones, concluding that the excessive use of mobile phones in university students are affected in their academic performance, as well as the relationship with patients and their colleagues at the time of carrying out their pre-professional practices.

Another study in Spain [14], argued that the excessive use of mobile phones taking the time of use, the age and the academic qualifications of the university students are factors that influence their self-esteem, therefore the misuse of mobile phones of students can generate addictive behaviors in the short and long term, compromising their health status and generating low self-esteem.

In Malaysia [15], a study was conducted in 119 students where they stated in their results that those who did not use mobile phones had a better recovery in terms of learning and memory, while those who used mobile phones showed that it affects considerably learning and memory of students.

The objective of the research work is to determine the problematic use of mobile phones during the COVID-19 pandemic in Peruvian university students, 2020. This study is important since it will provide us with relevant and real data about the mental vulnerability and addiction that university students have during the pandemic.

II. Methodology

In this part, the type and design of the research will be developed, as well as the population and sample that will be carried out in the research work, the inclusion criteria in detail and finally the data collection technique and instrument.

A. *Research Type and Design*

The present study, due to its characteristics, way of collecting data and measuring the variables involved, has a quantitative approach. Regarding the methodological design, it is a non-experimental, descriptive, cross-sectional study [16].

B. *Population and Sample*

In this research work, it will be made up of 163 university students from Lima, Piura, and Trujillo.

C. *Inclusion Criteria*

- University students who are studying from the 1st semester to the 10th semester.
- University students who agree to be voluntarily participate in the research work.

D. *Technique and Instrument*

The technique used is the virtual Google form survey, in which the objective of the Mobile Phone Problem Use Scale (MPPUS) data collection instrument is to measure the problematic use of mobile phones during the COVID-19 pandemic in Peruvian university students, 2020.

For data collection, it was structured in two blocks: 1) Sociodemographic data such as age, sex, marital status, type of family, university, career and semester of study; 2) Problematic use of mobile phones (MPPUS) that contains 27 items valued on a Likert-type scale where “1 = totally disagree”, “2 = disagree”, “3 = neither agree nor disagree”, “4 = agree” and “5 = totally agree”, the final score ranges from 27 to 135, where “27 to 50 is low problem use”, “51 to 74 is medium problem use” and “75 to 135 high problem use”, the higher the score corresponds to the more problematic use of mobile phones [17].

High problematic use refers to the excessive time or management that the university student has with the use of mobile phones, medium problematic use, refers to the time or habitual management that the university student has in the use of mobile phones and low problematic use, is refers to the time or occasional management that the university student has in the use of mobile phones during the COVID - 19 pandemic.

Regarding the validation and reliability results, they were obtained using the Kaiser-Meyer-Olkin sampling adequacy tests and the Bartlett sphericity test. The results of the sphericity test are significant while the sample adequacy test obtained a value of 0.538 (KMO > 0,6) and the results of the sphericity test obtained a favorable value (X^2 = 1570,572; g.l. = 351; p = 0,000). Likewise, Cronbach's alpha index obtained a value of 0.684 ($\alpha > 0,6$). All the above tests confirm an acceptable index of validity and reliability of the instrument.

E. *Place and Application of the Instrument*

The virtual survey was carried out to measure the problematic use of mobile phones due to the COVID-19 pandemic in university students from different regions of Peru, in which it was carried out in Universities of Lima, Piura and Trujillo.

In order to start the data collection process, it coordinated with university students from the Universidad de Ciencias y Humanidades, Universidad César Vallejo, Universidad Privada del Norte, Universidad de Piura, Universidad Nacional de Cajamarca, Universidad Técnologica del Perú and Universidad San Martin de Porres, to voluntarily participate in research work, although there were certain limitations since not all students were willing to participate in research work, for work reasons or other corresponding factors in virtual classes or at home.

Likewise, the virtual survey of Peruvian university students is based on modern technology that they predispose for their management and in what way they spend time on it, therefore, the way in which university students can adapt during the COVID-19 pandemic to the use of modern technology in mobile phones.

III. Results

Below is a summary table of the surveys carried out following the guidelines corresponding to the research work:

Fig. 1 shows the data of the university students in the study, where it can be observed with respect to the problematic use of mobile phones that 103 (63.2%) of the university students have a high problematic use, 59 (36.2%) medium problem use and 1 (0.6%) low problem use.

This is since many of the university students tend to exceed the prudent time of using the technology of mobile phones exceeding between 8 to 12 corresponding hours, due to this each time the use of mobile phones becomes more problematic and that can generate an addiction in it.

Fig. 2 relates to problematic use of mobile phones and the sex of university students, where 77 (66.4%) of the female sex have a high problematic use and 39 (33.6%) of the female sex have a medium problem use, in males 26 (63.2%) have high problem use, 20 (36.2%) have medium problem use and 1 (0.6%) have low problem use.

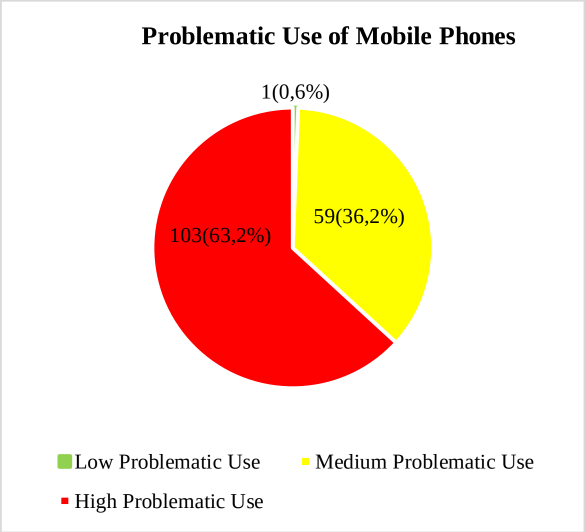


Fig. 1. Problematic use of Mobile Phones during the COVID-19 Pandemic in Peruvian University Students, 2020.

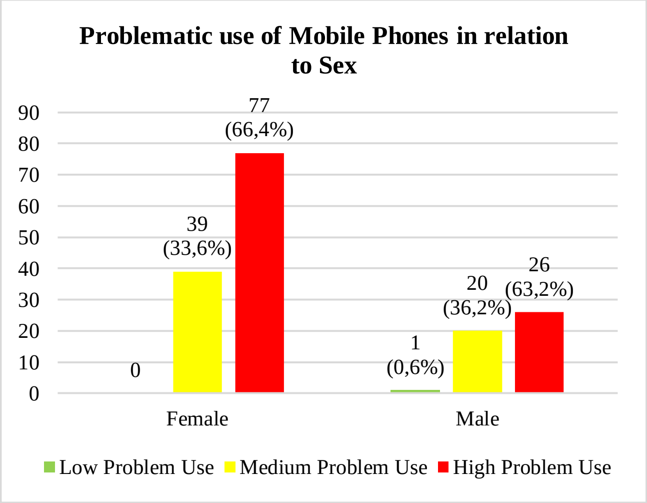


Fig. 2. Problematic use of Mobile Phones in relation to Sex during the COVID-19 Pandemic in Peruvian University Students, 2020

With respect to the female sex, a high index of problematic use of mobile phones was seen, this is because the female sex is more social than the male sex, because they are more aware of their emails or messages from friends that is why its use is considered extremely high.

Table I shows the problematic use of mobile telephones according to the university of Peruvian university students, in which it was determined with Pearson's chi-square test (X^2). The level of significance of the test obtained a value of 0.06 (p> 0.05) (X^2 = 2,958; d.f = 12). Therefore, an association hypothesis is not rejected, for which there are statistical data that verify the relationship between the problematic use of mobile phones and the University. By which, we can interpret that the university students of the Universidad de Ciencias y Humanidades have a high problematic use (62.5%), as in the Universidad de Piura (58.3%), the Universidad Privada del Norte (58, 1%), the Universidad Nacional de Cajamarca (60%), the Universidad Técnologica del Perú (66.7%), Universidad César Vallejo (71.4%) and the Universidad San Martín de Porres (69.2%).

In Table II, the problematic use of mobile phones is related according to the career of Peruvian university students, in which it was determined with Pearson's chi-square test (X^2). The level of significance of the test obtained a value of 0.06 (p>0.05) (X^2 = 4,364; d.f = 12). Therefore, an association hypothesis is not rejected, for which there are statistical data that verify the relationship between the problematic use of mobile phones and the career of university students. By which, it can interpret that the university students of the nursing career have a high problematic use (63.6%), as well as the Economics career (58.3%), the Administration career (60%), the Law (66.7%), Education (70%) and Accounting (72%), finally, university students in the Computer Systems Engineering career have a medium problematic use of mobile phones.

These results show us how university students handle the use of mobile phones during the COVID-19 pandemic, also will contribute to more future research work on the current research topic.

TABLE I. PROBLEMATIC USE OF MOBILE PHONES IN RELATION TO THE UNIVERSITY DURING THE COVID-19 PANDEMIC IN PERUVIAN UNIVERSITY STUDENTS, 2020

			Total Problematic Use			**Total**
			Low Problem Use	**Medium Problem Use**	**High Problem Use**	
University	**Universidad de Ciencias y Humanidades**	Count	1	23	40	64
		% within the University	1,6%	35,9%	62,5%	100,0%
	Universidad de Piura	Count	0	5	7	12
		% within the University	0,0%	41,7%	58,3%	100,0%
	Universidad Privada del Norte	Count	0	13	18	31
		% within the University	0,0%	41,9%	58,1%	100,0%
	Universidad Nacional de Cajamarca	Count	0	4	6	10
		% within the University	0,0%	40,0%	60,0%	100,0%
	Universidad Técnologica del Perú	Count	0	4	8	12
		% within the University	0,0%	33,3%	66,7%	100,0%
	Universidad César Vallejo	Count	0	6	15	21
		% within the University	0,0%	28,6%	71,4%	100,0%
	Universidad San Martín de Porres	Count	0	4	9	13
		% within the University	0,0%	30,8%	69,2%	100,0%
Total		Count	1	59	103	163
		% within the University	0,6%	36,2%	63,2%	100,0%

Chi-square tests			
	Value	df	Asymptotic significance (bilateral)
Pearson's Chi-square	2,958[a]	12	,996
Likelihood ratio	3,287	12	,993
N° of valid cases	163		

a. 11 cells (52.4%) have expected a count less than 5. The minimum expected count is .06.

TABLE II. PROBLEMATIC USE OF MOBILE PHONES IN RELATION TO THE RACE DURING THE COVID-19 PANDEMIC IN PERUVIAN UNIVERSITY STUDENTS, 2020

			Total Overuse			**Total**
			Low Problem Use	**Medium Problem Use**	**High Problem Use**	
Career	**Nursing**	Count	1	27	49	77
		% within the University	1,3%	35,1%	63,6%	100,0%
	Economy	Count	0	5	7	12
		% within the University	0,0%	41,7%	58,3%	100,0%
	Computer Systems Engineering	Count	0	9	8	17
		% within the University	0,0%	52,9%	47,1%	100,0%
	Administration	Count	0	4	6	10
		% within the University	0,0%	40,0%	60,0%	100,0%
	Law	Count	0	4	8	12
		% within the University	0,0%	33,3%	66,7%	100,0%
	Education	Count	0	3	7	10
		% within the University	0,0%	30,0%	70,0%	100,0%
	Accounting	Count	0	7	18	25
		% within the University	0,0%	28,0%	72,0%	100,0%
Total		Count	1	59	103	163
		% within the University	0,6%	36,2%	63,2%	100,0%

Chi-square tests			
	Value	df	Asymptotic significance (bilateral)
Pearson's Chi-square	4,364[a]	12	,976
Likelihood ratio	4,676	12	,968
N° of valid cases	163		

a. 11 cells (52.4%) have expected a count less than 5. The minimum expected count is .06.

IV. Discussion

In this research work, a focus was provided on the mental health of university students in relation to the problematic use of mobile phones during their isolation due to the COVID-19 pandemic.

In the results on the problematic use of mobile phones, we can see that Peruvian university students present because of the COVID-19 pandemic, it has generated consequences in college students, such a degree that their psycho-emotional health has been affected, giving college students such as boredom, personality change, anxiety, and depression to such an extent that the feeling of having their mobile phone can help them manage their psycho-emotional state. In the study by X. Yang et al. [1], they argue that high levels of problematic mobile phone use are more related to boredom in college students, it generates undesirable psychosocial consequences in college students, such as anxiety and depression, impaired cognition, and poor academic performance.

In the results in relation to sex, we can observe that the female sex presented a high level of problematic use of mobile phones, this is because women are more attentive to their mobile phones since, they use it for their social networks, joint activities with friends, or are on the screen of their phones while talking by call with a person. In the study by J. Gao and collaborators [5], they argue that women are the ones who use social networks more on their mobile phones than men, because social networks are the only method to reduce their anxiety and depression , the COVID-19 pandemic is expected to affect the female population more than the male population.

In relation to the career, it was found that nursing students present a high level of problematic use of mobile phones, this is due to the fact that students, because of the COVID-19 pandemic, have had to stop their professional practices already also, take virtual courses either through their computers or mobile phones, generating stress and anxiety in themselves since they are not used to using technology on a daily basis since, as health students, they are more oriented to care and not to technology. In the study by V. Márquez et al. [13], they argue that female nursing students have a greater tendency to use mobile phones more than males, since females engage and maintain more active interpersonal relationships than males.

These aspects mainly encompass in mental health what K. Chung et al. [6], maintain, referring that the inappropriate use of mobile phones compromises both the mental and physical health of the person, relating it to depressive symptoms, body pain and daytime sleepiness as predisposing factor. O. Király and collaborators [18], argue that the problematic use of mobile phones due to the COVID-19 pandemic has generated that the levels of uncertainty about the future and the insecurity, also generate a high level of stress, anxiety and depression because of this, therefore using mobile phones helps reduce those psycho-emotional factors that affect mental health due to confinement and quarantine.

V. Conclusion

It is concluded that programs on mental health should be carried out during the COVID-19 pandemic in university students.

It is concluded that the use of mobile phones should not be over-limited since it is compromised in the professional training of university students.

It is concluded that the use of technology for nursing students is detrimental in their academic performance, in decision-making, in the nurse-patient relationship and in the relationship with other colleagues from the career.

It is recommended that university students carry out a schedule of activities each day of the week, to allow them to carry out their daily routines at home during isolation or quarantine.

It is recommended to sleep regularly, have a healthy diet, take care of their personal hygiene, since they are essential to maintain good physical health and allows to improve their psychological well-being.

The limitation in this research work is that there are not many studies on the subject that are being carried out, both nationally and internationally.

References

[1] X. Yang, Q. Liu, S. Lian, and Z. Zhou, “Are bored minds more likely to be addicted? The relationship between boredom proneness and problematic mobile phone use.,” Addict. Behav., vol. 108, p. 106426, 2019, doi: 10.1016/j.addbeh.2020.106426.

[2] Z. Vally and F. El Hichami, “An examination of problematic mobile phone use in the United Arab Emirates: Prevalence, correlates, and predictors in a college-aged sample of young adults.,” Addict. Behav. Reports, vol. 9, p. 100185, 2019, doi: 10.1016/j.abrep.2019.100185.

[3] G. Zhang, X. Yang, X. Tu, N. Ding, and J. Lau, “Prospective relationships between mobile phone dependence and mental health status among Chinese undergraduate students with college adjustment as a mediator.,” J. Affect. Disord., vol. 260, pp. 498–505, 2020, doi: 10.1016/j.jad.2019.09.047.

[4] Hong W, Liu R, Ding Y, Sheng X, and Zhen R, “Mobile phone addiction and cognitive failures in daily life: The mediating roles of sleep duration and quality and the moderating role of trait self-regulation.,” Addict. Behav., vol. 107, no. 19, p. 10638, 2020, doi: 10.1016/j.addbeh.2020.106383.

[5] J. Gao et al., “Mental health problems and social media exposure during COVID-19 outbreak.,” PLoS One, vol. 15, no. 4, pp. 1–10, 2020, doi: 10.1371/journal.pone.0231924.

[6] K. Chun et al., “The relationships between mobile phone use and depressive symptoms, bodily pain, and daytime sleepiness in Hong Kong secondary school students.,” Addict. Behav., vol. 101, p. 105975, 2020, doi: 10.1016/j.addbeh.2019.04.033.

[7] M. Romano, L. Osborne, R. Truzoli, and P. Reed, “Differential Psychological Impact of Internet Exposure on Internet Addicts.,” PLoS One, vol. 8, no. 2, pp. 8–11, 2013, doi: 10.1371/journal.pone.0055162.

[8] H. Dong, F. Yang, X. Lu, and W. Hao, “Internet Addiction and Related Psychological Factors Among Children and Adolescents in China During the Coronavirus Disease 2019 (COVID-19) Epidemic,” Front. Psychiatry, vol. 11, p. 751, 2020, doi: 10.3389/fpsyt.2020.00751.

[9] A. Ruiz, M. López, and J. López, “Evitación experiencial y uso abusivo del smartphone: un enfoque bayesiano,” Adicciones, vol. 32, no. 2, p. 116, 2020, doi: 10.20882/adicciones.1151.

[10] S. Amez and S. Baert, “Smartphone use and academic performance: A literature review,” Int. J. Educ. Res., vol. 103, p. 101618, 2020, doi: 10.1016/j.ijer.2020.101618.

[11] S. Lian, X. Sun, G. Niu, X. Yang, Z. Zhou, and C. Yang, “Mobile phone addiction and psychological distress among Chinese adolescents: The mediating role of rumination and moderating role of the capacity to be alone.,” J. Affect. Disord., 2020, doi: 10.1016/j.jad.2020.10.005.

[12] Z. Hao et al., “Alexithymia and mobile phone addiction in Chinese undergraduate students: The roles of mobile phone use patterns.,” Comput. Human Behav., vol. 97, pp. 51–59, 2019, doi: 10.1016/j.chb.2019.03.001.

[13] V. Márquez, L. Puertas, G. Gámez, V. Puertas, and G. Manrique, “Problematic mobile phone use, nomophobia and decision-makin in nursing students mobile and decision-makin in nursing students.,” Nurse Educ. Pract., vol. 49, p. 102910, 2020, doi: 10.1016/j.nepr.2020.102910.

[14] J. Romero and I. Aznar, “Análisis de la adicción al smartphone en estudiantes universitarios. Factores influyentes y correlación con la autoestima,” Rev. Educ. a Distancia, vol. 60, no. 8, 2019, [Online]. Available: https://revistas.um.es/red/article/view/396051/273191.

[15] C. Tanil and M. Yong, “Mobile phones: The effect of its presence on learning and memory,” PLoS One, vol. 15, no. 8, pp. 1–12, 2020, doi: 10.1371/journal.pone.0219233.

[16] C. Fernández and P. Baptista, Metodología de la Investigación. 6ta ed. México: Mc Graw-Hill/Interamericana. 2015.

[17] O. López, M. Honrubia, and M. Freixa, “Adaptación española del ‘Mobile Phone Problem Use Scale’ para población adolescente.,” Adicciones, vol. 24, no. 2, p. 123, 2012, doi: 10.20882/adicciones.104.

[18] O. Király et al., “Preventing problematic internet use during the COVID-19 pandemic: Consensus guidance,” Compr. Psychiatry, vol. 100, p. 152180, 2020, doi: 10.1016/j.comppsych.2020.152180.

Examining Users' Willingness to Post Sensitive Personal Data on Social Media

O'la Hmoud Al-laymoun[1], Ali Aljaafreh[2]
Mutah University
Alkrak, Jordan

***Abstract*—Reaping the vast benefits of ubiquitous social media requires users to share their information, preferences, and interests on these websites. This research draws on communications privacy management theory and the online privacy literature to develop and validate an empirical research model testing users' willingness to share sensitive data on Facebook. The data were collected using an online survey from 569 respondents, however; 515 responses were valid for the statistical analysis. The valid data were analyzed using SMART-PLS2. The findings showed the need for attention as a significant predictor of Facebook users' willingness. Neither individual perceptions of privacy control nor privacy risks had an impact on the variable of interest. Furthermore, the evidence supported the positive impact of each of deposition to value privacy and the perceived effectiveness of Facebook's privacy policy on mitigating Facebook users' perceptions of the risks of posting their private data on the website. The paper discusses the study's theoretical and managerial implications along with its limitations.**

Keywords—Self-disclosure; sensitive data; Facebook policy; government regulations; privacy control; privacy risk

I. Introduction

Facebook is the largest social network, with about 2.5 billion monthly active worldwide users as of December 31, 2019 [1]. It is not surprising that increasing numbers of people are joining Facebook, as it offers its users a wide range of benefits. According to Statista (2019), which examined the main reasons for using Facebook among 2,100 U.S citizens 15 years or older, 88% of participants reported staying in touch with family and friends as the top reason, followed by getting entertainment (33%), getting news (23%), following brands and companies (17%), and strengthening their professional networks (11%).

Unfortunately, there are risks to reaping the benefits of Facebook, as it requires its users to share information with others, creating a priceless treasure of personally identifiable information for businesses and cybercriminals to exploit [2]. For example, according to a 2018 report by Forbes.com, more than 300 million photos are uploaded to Facebook daily, and about 510,000 comments and 293,000 status updates are posted on the website every minute. Therefore, it is not surprising that Facebook represents a precious target for cybercrimes, such as privacy violations. Privacy protection is especially challenging in the era of social networking, as the world does not have or enforce the right laws and regulations to deal with a rapidly changing digital environment [3]. Privacy protection challenges are pushing lawmakers to rely more on today's empowered consumers to make the right decisions to protect themselves and their privacy while online [3]. Examples of online self-protection behaviors include managing privacy preferences and sharing information with websites that promise not to share that information with third parties [4].

Online privacy and self-disclosure on social networking sites (SNSs) and other websites have received a good deal of researchers' attention [2], [3], [5]–[11]. Yet, little has been done to understand individuals' willingness to put their sensitive information online, except by Widjaja et al. [12], which has been applied to the context of cloud storage. SNSs represent a fertile environment for this kind of research due to the increasing numbers of subscribers and the diverse potential sources of privacy threats from the website itself, as when Facebook gave Cambridge Analytica access to the data of 50 million users [13], other users, governments, or businesses. Thus, this research paper represents an attempt to fill in this gap in the online privacy literature by concentrating on the self-disclosure of sensitive data on SNSs, especially Facebook. Furthermore, this study took place in Jordan, a Middle Eastern country in which about 70% of the population uses Facebook [14]. Yet there is a dearth of research investigating Jordanian users' online privacy-related behavior. As a result, this study explores this understudied context. Specifically, this paper addresses the following research question: What factors influence Facebook users' willingness to put their sensitive information on Facebook?

II. Literature Review

A. Communication Privacy Management

This research is based on the communication privacy management theory (CPMT) proposed by Petronio [15] to study information disclosure in interpersonal relationships. Researchers have applied CPMT to understand relationships within groups, organizations, and institutions in online and face-to-face contexts [4]. CPMT has three main premises: boundary rule formation, boundary coordination, and boundary turbulence. According to CPMT, information disclosure to others has potential risks, as it makes one vulnerable to exploitation by others [15]. Nonetheless, nondisclosure has its drawbacks, as it deprives one of the benefits of disclosure, including making friends and receiving social support. Thus, when interacting with others, one goes through a risk-control assessment to weigh the rewards of disclosure and the level of control one has over the revealed information against the potential risks [16]. Based on that privacy calculus and other personality and environment-related factors, one draws a hypothetical boundary specifying one's private space [16]. Such boundaries are regulated by rules that manage information flow in and out of the private informational space through opening (disclosing information) and closing (withholding information) boundaries. Accordingly, boundary management reflects one's perception of privacy and serves as a means of self-protection. People who impose strict control on their boundaries, by limiting revealing information to others, lower their vulnerability to exploitation and privacy concerns and vice versa [15], [17].

Furthermore, CPMT emphasizes that individuals are the owners and, as such, need to keep control of their private information even after voluntarily sharing with others who become co-owners in that case. This partnership creates a need for boundary coordination among both parties, which refers to agreeing upon privacy rules that meet the privacy expectations of the owners of the information [4]. In the case of boundary miscoordination or privacy rules violations, boundary turbulence occurs [4]. In this case, individuals seek the help of third parties by, for example, filing complaints [17].

B. Willingness to Share Sensitive Information on Facebook

The current research paper uses CPMT to investigate the role of many personal and environmental factors that motivate or hinder people from posting their private information on Facebook. The status update box on Facebook induces users to share their thoughts, media, and almost all kinds of personal information readily with their friends or even with the public [6]. According to Widjaja et al [12], there are five types of personal information representing different information sensitivity levels. These are from the least to the most sensitive: "work-related documents, personal media, personal documents, personal identity information, and specific sensitive information."

As information sensitivity increases, so does the risk associated with disclosure, making the necessary boundary management stricter [15] and lowering information disclosure [18], [19]. Research has shown that consumers' privacy concerns and willingness to share their personal information with marketers depend on information type, such that people are least open to reveal financial and personal identifier information and more willing to provide demographic and lifestyle-related information [20]. Patil and Kobsa [21] pointed out that one of the main reasons for instant messaging users' privacy concerns is content sensitivity. Metzger[4] also found that online consumers tend to protect their privacy online by withholding or falsifying sensitive information. Metzger found that consumers were most likely to withhold their financial information and information that could be linked to it, such as their social security numbers, personal contact information, and preferences, whereas they were more open to disclose their demographic information.

III. Hypothesis Development

Perceived privacy risk (PPR) is defined as the anticipated losses resulting from online disclosure of private information [16]. It is a negative belief that is expected to influence one's privacy concerns [17] as it stems from the potential risk of being a victim of opportunistic behavior [22] resulting from information misuse [17]. Generally speaking, people weigh the benefits and risks of disclosure when participating in social exchange situations and, as a result, choose to disclose if the benefits outweigh the risks or withhold if the opposite is true [5]. Many studies have examined the consequences of PPR. For example, it has had a negative association with consumers' willingness to transact online [23]. In the same context, Dinev and Hart [22] found it had a negative relationship with willingness to provide personal information to participate in online transactions. In addition, Millham and Atkin [10] found that the higher one's perception of the risks resulting from disclosing information on online social networks, the greater the sense of information ownership and responsibility, which, in turn, leads to lower willingness to reveal sensitive personal information on these networks. In line with the previous studies, we posit the following:

H1: Perceived privacy risk has a negative impact on users' willingness to post sensitive personal information on Facebook.

Perceived privacy control (PPC) is defined as one's perception of the ability to control the collection, dissemination, and the resulting use of one's private information [16]. In general, when individuals do not have that control or when they are not aware of the subsequent uses of their information, they tend not to disclose [24]. According to CMPT, individuals are the owners of their personal information and, thus, they need to keep it under their control [12]. Indeed, a recent study has shown that most online consumers are worried about how companies handle their information and seek more control over the ways businesses process that information [20] to mitigate the possible risks of disclosure [24]. In the context of e-commerce, Phelps et al. [20] found that consumers' information control had a positive association with online shopping intention; however, the researchers also reported that intention was higher when consumers were asked to submit lifestyle or demographic information than when they were asked to submit financial or personal identifier information. In another study by Benson et al. [9], users' control over personal data had a negative relationship with information disclosure in the context of SNSs. On Facebook, the privacy and security threats are not limited to the misuse of published content by the website. One's network friends and even strangers could also be sources of all kinds of violations, making publishing one's private data a sensitive matter. Indeed, Zlatolas et al. [8] found that privacy control had no significant impact on the self-disclosure behavior of Facebook users. Thus, based on the aforementioned studies, we propose the following:

H2: Perceived privacy control has no impact on users' willingness to post sensitive personal information on Facebook.

We are extending CMP theory by adding the need for attention construct, which can impact online user behavior, such as posting malicious comments [25] and online political content generation and consumption [26]. In the context of social media, users face information overload as they receive all kinds of digital content from their online friends and from strangers. Thus, they have to be selective in the content they pay attention to and interact with. To stand out from the crowd and increase the attention their posts receive from others, people with a high need for attention might tend to have a high frequency of online content creation and to select content that is likely to attract others. We expect posting sensitive data on Facebook to serve that goal by providing material unique to the individual instead of presenting general information. Indeed, in a longitudinal study, Hawk et al. [27] found that adolescents' attention-seeking motives had a positive relationship with their self-disclosure on social media. Thus, we propose the following:

H3: Need for attention has a positive impact on users' willingness to post sensitive personal information on Facebook.

Disposition To Value Privacy (DTVP) is an inherent personal need to maintain one's private space and control the flow of information outside that space [16], [17]. DTVP is analogous to the privacy orientation construct in CPMT [12]. According to Widjaja et al.[12], people who score high on DTVP value privacy more and perceive higher privacy and security risks in disclosure than those who score low on that construct. For example, Patil and Kobsa [21] found a relationship between DTVP and instant messaging users' privacy concerns. The positive relationship between DTVP and PPR has also found support in several other contexts, such as cloud-based storage applications [12], e-commerce, SNSs, financial, and healthcare sites [16], [17]. Thus, in line with the extant literature, we postulate the following:

H4: Disposition to value privacy has a positive impact on perceived privacy risk.

The online environment is risky, and information disclosed online could be misused by, for example, being sold to third parties without consent. Once individuals provide their information to a website, it becomes hard for them to remove it or even to control its subsequent use [28]. Information asymmetries increase the complexity of that situation by limiting individuals' awareness of the organization's information practice and whether their collected information may be misused [29]. Per CMPT, once one shares one's information with others, they become co-owners. Both parties need to negotiate the owners' privacy expectations regarding how the co-owners will use and handle that information and who else can access it in a process called boundary coordination [30]. In online contexts, a website could address its customers' privacy concerns and signal that it is a trustworthy custodian of their information by using institutional privacy assurances, defined as interventions taken by the company to protect and keep the privacy and safety of its customers' information[15] [30], such as privacy policies [16] and notices [29].

Privacy policies are mechanisms informing individuals of the subsequent uses of their information, the safety measures and privacy rules used to protect their information from different kinds of misuse, and the ways available to them to keep their information accurate and up-to-date [16]. They communicate whether, how, and when consumers' private information will flow out the collective boundary after being disclosed, helping users to decide whether their acceptable privacy rules and the organization's rules align [4] and enhancing users' overall regard for and trust of the organization [5]. In the context of e-commerce, Jarvenpaa et al. [23] found that the higher consumers' trust in a website, the lower the perceived risk of purchasing from that website. Chellappa and Sin [31] pointed out that individuals' usage of online personalization services had a positive relationship with their trust in the online merchant; thus, they suggested that vendors need to use trust-building methods and tools if they want to collect and capitalize on their consumer's data. Furthermore, in some cases, when users were informed that fair information practices are in place to protect their information, privacy concerns did not differentiate those who were willing to have their information used for profiling from those who were not [32]. Interestingly, in an experimental setting of e-commerce, Jensen et al. [33] found that the existence of privacy policies impacted participants' behavior, although they were rarely consulted. Thus, we propose the following hypothesis:

H5: The perceived effectiveness of Facebook privacy policy a) negatively impacts perceived privacy risks and b) positively impacts perceived privacy control.

Sometimes, privacy policies and other forms of institutional privacy assurances used by organizations to assure their customers that their information will be kept confidential and safe are not adequate to meet those customers' expectations. According to CPMT, when boundary coordination fails, boundary turbulence occurs. Boundary turbulence results from privacy violations, lack of boundary coordination, or conflicting privacy rules used by different people [5]. In that case, consumers tend to turn to other forms of institutional privacy assurances, such as industry self-regulation and government regulations, to protect their privacy [12].

The literature on privacy has emphasized government legislation as one of the main approaches individuals use to maintain their online and offline privacy [12]. Like other countries, Jordan has special legislation in place to combat online crimes. The Cybercrimes Unit in Jordan's Public Security Department is the official party that Jordanians turn to if they become victims of cybercrimes. Widjaja et al. [12] found that the perceived effectiveness of government regulations in enhancing users' perceived privacy control had no impact on the perceived cost of putting sensitive data on cloud-based applications. However, in line with CPMT, we expect it to have a positive influence on PPC and a negative association with PPR. Thus, we propose the following:

H6: The perceived effectiveness of government regulations a) negatively impacts perceived privacy risks and b) positively impacts perceived privacy control.

IV. Research Methodology

A. Data Collection and Instrument Development

A survey was employed to collect the primary data from Facebook users in a large public university in Jordan. The data were collected using a questionnaire developed on Google Forms and attached with a cover letter, assuring the confidentiality of research participants and outlining the study's primary purpose. A pilot study was conducted with seven undergraduate students, and, as a result, minor modifications were made to the initial version of the questionnaire. The questionnaire was posted to 32 teams on Microsoft Teams, corresponding to 32 different classes taught on that application during the lockdown in Jordan during the “COVID-19 pandemic”. Professors encouraged their students to participate in the study. Participation was voluntary, and no incentives were available to the research subjects. The data collection process took about a month and a half. The total number of received responses was 526, of which 515 were retained for further analysis while the rest were dropped from the study due to inconsistent answers. The demographic characteristics of the research sample are outlined in Table I.

TABLE I. DEMOGRAPHIC CHARACTERISTICS OF THE SAMPLE

Measure	Item	Frequency	(%)
Gender	Male	153	29.7
	Female	362	70.3
Facebook Daily hours	Less than 1 hour	128	24.9
	1 to less than 2 hours	90	17.5
	2 to less than 3 hours	97	18.8
	3 to less than 4 hours	72	14.0
	4 to less than 5 hours	41	8.0
	More than 5 hours	87	16.9
Age	18-22	249	48.3
	23.26	126	24.5
	27-30	39	7.6
	31-34	23	4.5
	35-38	28	5.4
	39-42	34	6.6
	More than 42	16	3.1
Education	Bachelor’s	401	77.9
	Master’s	87	16.9
	PhD	27	5.2

B. Measures

A questionnaire with a 5-point Likert-type scale (ranging from 1 = “strongly disagree” to 5 = “strongly agree”) was employed to collect the data. All measures were adopted from previous studies and adapted as needed to achieve the purpose of this research. Willingness to post sensitive personal data on Facebook measures Facebook users’ readiness to make their private data available for other users on the website. The construct adopted from [12] asked the research subjects to indicate their willingness level to put five different types of personal information on the social network. Perceived privacy risk measures one’s cognitive assessment of the potential privacy threats associated with personal data availability on Facebook. The variable was adapted from [17] and consisted of four items. Perceived privacy control reflects one’s evaluation of the ability to manage what to post on Facebook, who can view that content, and controlling how Facebook can use that data. The variable was measured using four items, and it was adopted from [17]. The need for attention reflects one’s desire to be noticed and appreciated by others. Five items adopted from [25] were used in this study. However, one of the items (“I don’t like people who do not respond to my post on Facebook”) was dropped for not loading well on the latent variable. Disposition to value privacy was adopted from [17]. This variable used three items to measure the predisposition to value privacy online and offline; however, one item (“Compared to others, I tend to be more concerned about threats to my information privacy”) was dropped as it did not load well on the construct. The perceived effectiveness of Facebook policy refers to individuals’ evaluation of Facebook’s commitment and ability to protect its users' privacy. The construct was measured using three items adopted from [12]. Finally, the perceived effectiveness of government regulations is defined as the perception of the research subjects of the ability of the Cybercrimes Unit in Jordan to handle any privacy violation incidents they might face. Three items were adopted from [12], and one item developed by the researchers was used to measure the variable.

V. RESULTS

A. Measurement Model

The measurement model was evaluated using the convergent validity and discriminant validity of the survey. The research instrument’s reliability was assessed using two measures: the composite reliability (CR) and Cronbach’s alpha. The recommended threshold for both measures is 0.7 or above; however, a value of 0.5 is considered the minimum acceptable value [34]. According to the results presented in Table II, the composite reliability ranges between 0.868 and 0.943 substantially exceed the recommended criterion. In regard to Cronbach’s alpha, the values for the seven constructs were between 0.698 and 0.92. Based on these results, we feel confident in the high internal consistency of the research instrument. The research validity was tested using discriminant validity and convergent validity [35]. Average variance extracted (AVE) and factor loadings were employed to assess the convergent validity. The requirement of having an AVE of 0.5 or more has been satisfied, as Table II shows. Furthermore, all factor loadings exceeded the desired threshold of 0. 7. To ensure discriminant validity, each indicator’s factor loading should be higher on the factor it measures than any other factor. This condition was also met. Thus, convergent validity and discriminant validity have been established.

B. Structural Model

The findings of the PLS-SEM analysis are summarized in Table III and Fig. 1. The results indicate that three out of the seven hypotheses were statistically accepted. The analysis results showed that Disposition to value privacy (β =0.456, t-value=4.1944) had a significant positive impact on Perceived Privacy Risk, therefore; H1 has been confirmed. Also, the results showed that Need for Attention strongly affected Willingness to post sensitive data on Facebook (β =0.299, t-value=3.346), thus, the results confirmed the positive impact hypothesized in H4. Finally, the SEM analysis revealed that Facebook Policy was a significant predictor for Privacy Control (β =0.446, t-value=4.440), thereby supporting H8.

TABLE II. OVERVIEW OF QUALITY CRITERIA AND FACTOR LOADINGS

Construct	Items	Factor loadings	AVE	CR	Cronbach α
The perceived effectiveness of government regulations	CCU1	0.8865	0.806	0.943	0.920
	CCU2	0.8734			
	CCU3	0.92			
	CCU4	0.9119			
Disposition to value privacy	DP1	0.8717	0.768	0.868	0.698
	DP2	0.8812			
The perceived effectiveness of Facebook privacy policy	FBPE1	0.8468	0.774	0.911	0.854
	FBPE2	0.9081			
	FBPE3	0.8845			
Need for attention	NA1	0.8084	0.708	0.906	0.862
	NA2	0.8896			
	NA3	0.881			
	NA4	0.784			
Perceived privacy control	PC1	0.7889	0.701	0.802	0.858
	PC2	0.8479			
	PC3	0.8762			
	PC4	0.8352			
Perceived privacy risk	PR1	0.8372	0.671	0.890	0.838
	PR2	0.8787			
	PR3	0.7358			
	PR4	0.8203			
Willingness to post sensitive personal data	W1	0.7188	0.504	0.802	0.678
	W2	0.7106			
	W3	0.703			
	W4	0.7085			

TABLE III. PLS COEFFICIENT PATH ANALYSIS

Hypotheses	Beta (β)	t-value	results
H1. Perceived privacy risk -> Willingness to post	0.167	1.4444	rejected
H2. Privacy Control -> Willingness to post	0.071	0.6399	rejected
H3. Need for Attention -> Willingness to post	0.299	3.3468	Accepted
H4. Disposition to privacy -> Perceived privacy risk	0.456	4.1944	Accepted
H5.a. Facebook policy -> Perceived privacy risk	0.078	0.7231	rejected
H5.b. Facebook policy -> Privacy Control	0.446	4.4404	Accepted
H6.a. Government regulations -> Perceived privacy risk	0.058	0.5661	rejected
H6.b. Government regulations -> Privacy Control	0.190	1.56	rejected
H6.a. Government regulations -> Perceived privacy risk	0.058	0.5661	rejected
H6.b. Government regulations -> Privacy Control	0.190	1.56	rejected

[a] Note: Explained variance proportion R^2 of Willingness to post = 0.118, Explained variance proportion R^2 of Perceived privacy risk = 0.214, Explained variance proportion R^2 of Privacy Control = 0.318.

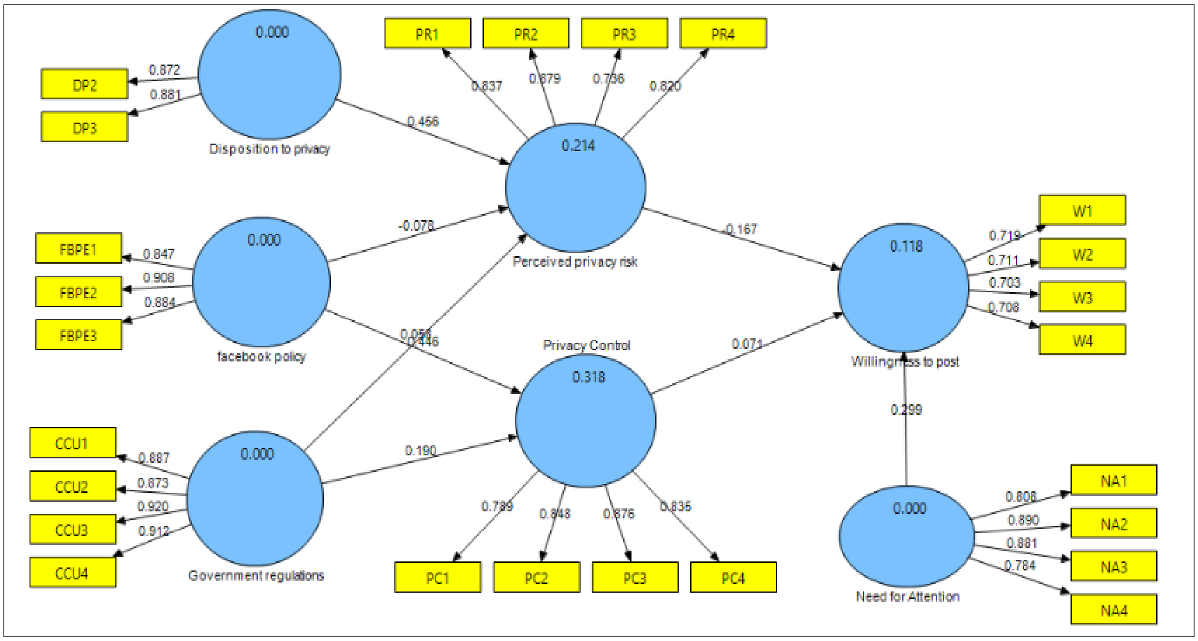


Fig. 1. PLS Path Analysis.

VI. Discussion

This study is an attempt to add to the literature on online privacy management. Specifically, we are looking at Jordanian Facebook users' willingness to post their sensitive personal data on Facebook. The study leverages the extant literature on privacy to develop and to empirically examine a research model combining CPMT and the need for attention construct to understand better what motivates or hinders users from revealing their private data on SNSs.

The results showed that the proposed model accounts for about 11.8% of the dependent variable, indicating that more investigation is still needed in this regard. In general, the results provided reasonable evidence of the role of the psychological need for attention as a motivator feeding individuals' readiness to share their sensitive data with their Facebook friends and even with strangers. This result signals the crucial influence of one's inner needs in shaping one's acceptance of specific behaviors. In line with our expectations, the study revealed no impact on perceived privacy control on the willingness to post private data on Facebook. Although this result confirms the results of some earlier studies [8], it contradicts other studies that either found positive or negative relationships between the aforementioned variables. This finding suggests that further research is needed to clarify the nature of the relationship between the variables and what moderators, if any, influence it. Furthermore, contrary to our expectations, perceived privacy risks had no significant relationship with individuals' willingness to share. This result is consistent with the findings of [36]. This may in part be attributable to the culture. Although people cognitively assess the likely risks of online self-disclosure and their control over their data, they might perceive the rewards of doing so as overweighing the risks. For instance, in a collectivistic culture like Jordan, people are more prone to social influence than in individualistic cultures. Social influence can impact the intention of self-disclosure on social media positively [36].

Moreover, the study found a positive influence for DTVP on PPR. In terms of institutional privacy assurances, two forms were investigated in this paper: the effectiveness of Facebook policy and the effectiveness of government privacy regulations. Consistent with the previous studies in this research area, the empirical evidence found that Facebook policy enhanced Facebook users' sense of control over their data posted on Facebook. However, our study found no significant impact of that policy on the perceptions of risks. These findings imply that policies play a role in assuring users that they are the owners of their data and that Facebook empowers them to manage it, yet these policies are running behind in terms of addressing and educating people about the potential vulnerabilities of being victims of privacy violations. With regard to the perceived effectiveness of government regulations, we found no support for its impact on PPC or PPR. Our findings are partially consistent with the study by [12]. They found no impact on the perceived effectiveness of government regulations in the context of cloud-based storage applications on PPR. However, it contradicts the evidence reported in the literature on the impact on PPC. This may be because the laws and regulations we have today are still not adequately addressing the fast-changing and very diverse online security and privacy violation domains [3]. People might doubt governments' ability to give them complete control over their online information and its subsequent uses. They also might question how effective these regulations would be.

The empirical evidence from the study has several important implications for researchers and practitioners alike. In terms of practical implications, finding a significant impact of the need for attention on people's willingness to put their sensitive data on Facebook highlights the importance of understanding social media users' psychological needs. The data online surfers post on SNSs represents a priceless treasure that businesses, governments, and other parties can mine to understand and target their audiences better. Thus, investing in big data, data mining, and other technologies to understand online users becomes a necessity. Moreover, Facebook needs to continue improving its filtering, recommendations, privacy management, and other tools to create safe and secure social environments that induce people to network and make friendships with others without being afraid of privacy and security threats. The study also makes theoretical contributions. First of all, while most research on online self-disclosure pays little attention to information sensitivity [12], especially when examining social media, this study takes that important matter into consideration. As indicated in previous research, people were more likely to self-disclose when they were asked to reveal low sensitivity information. In addition, to the best of the researcher's knowledge, the current study is one of the few empirical studies that has applied CPMT to investigate online user behavior in Jordan and the Arab world in general.

VII. LIMITATIONS AND FUTURE RESEARCH

Generally speaking, research studies can suffer from different kinds of drawbacks. This paper is no exception. First of all, users' perceptions of risk and, in turn, their willingness to reveal private data on Facebook could be a function of whether their account is private or public. No differentiation between account types has been made in this study. This factor could be studied in future research. Second, although this study focused on Facebook, its main premises could be extended to other social media applications. Third, it would be interesting to examine the research model in different cultures and to measure individuals' perceived willingness to post and their actual behavior.

VIII. CONCLUSION

Drawing on the communications privacy management theory and the online privacy literature, we developed a research model investigating users' willingness to share sensitive data on Facebook. A survey was used to collect the data from Facebook users in Jordan. The posited model explains about 11.8% of users' willingness to post personal data on the network. The results showed the need for attention as a significant predictor of Facebook users' willingness, whereas neither individual perceptions of privacy control nor privacy risks had a significant impact. The preliminary empirical evidence from this study sheds light on the importance of the psychological needs in shaping one's online behavior. It also opens the doors for future research to explore this novel area of research.

REFERENCES

[1] Facebook, "Facebook Reports Fourth Quarter and Full Year 2012 Results," 2020.

[2] A. Vishwanath, W. Xu, and Z. Ngoh, How people protect their privacy on facebook: A cost-benefit view, vol. 69, no. 5. 2018.

[3] S. C. Boerman, S. Kruikemeier, and F. J. Zuiderveen Borgesius, "Exploring motivations for online privacy protection behavior: Insights from panel data," Communic. Res., p. 0093650218800915, 2018.

[4] M. J. Metzger, "Communication privacy management in electronic commerce," J. Comput. Commun., vol. 12, no. 2, pp. 335–361, 2007.

[5] M. J. Metzger, "Privacy, trust, and disclosure: Exploring barriers to electronic commerce," J. Comput. Commun., vol. 9, no. 4, p. JCMC942, 2004.

[6] E. E. Hollenbaugh and A. L. Ferris, "Facebook self-disclosure: Examining the role of traits, social cohesion, and motives," Comput. Human Behav., vol. 30, pp. 50–58, 2014.

[7] A. Gruzd and Á. Hernández-García, "Privacy concerns and self-disclosure in private and public uses of social media," Cyberpsychology, Behav. Soc. Netw., vol. 21, no. 7, pp. 418–428, 2018.

[8] L. N. Zlatolas, T. Welzer, M. Heričko, and M. Hölbl, "Privacy antecedents for SNS self-disclosure: The case of Facebook," Comput. Human Behav., vol. 45, pp. 158–167, 2015.

[9] V. Benson, G. Saridakis, and H. Tennakoon, "Information disclosure of social media users," Inf. Technol. People, 2015.

[10] M. H. Millham and D. Atkin, "Managing the virtual boundaries: Online social networks, disclosure, and privacy behaviors," New Media Soc., vol. 20, no. 1, pp. 50–67, 2018.

[11] L. Yu, H. Li, W. He, F.-K. Wang, and S. Jiao, "A meta-analysis to explore privacy cognition and information disclosure of internet users," Int. J. Inf. Manage., vol. 51, p. 102015, 2020.

[12] A. E. Widjaja, J. V. Chen, B. M. Sukoco, and Q.-A. Ha, "Understanding users' willingness to put their personal information on the personal cloud-based storage applications: An empirical study," Comput. Human Behav., vol. 91, pp. 167–185, 2019.

[13] K. Wagner, "Here's how Facebook allowed Cambridge Analytica to get data for 50 million users - Vox," Mar. 2018.

[14] Statista, "• Jordan: share of Facebook users 2017 | Statista," 2017.

[15] S. Petronio, "Communication boundary management: A theoretical model of managing disclosure of private information between marital couples," Commun. theory, vol. 1, no. 4, pp. 311–335, 1991.

[16] H. Xu, T. Dinev, H. J. Smith, and P. Hart, "Examining the formation of individual's privacy concerns: Toward an integrative view," ICIS 2008 Proc., p. 6, 2008.

[17] H. Xu, T. Dinev, J. Smith, and P. Hart, "Information privacy concerns: Linking individual perceptions with institutional privacy assurances," J. Assoc. Inf. Syst., vol. 12, no. 12, p. 1, 2011.

[18] P.-C. Sun, R. J. Tsai, G. Finger, Y.-Y. Chen, and D. Yeh, "What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction," Comput. Educ., vol. 50, no. 4, pp. 1183–1202, 2008.

[19] S. Yang and K. Wang, "The influence of information sensitivity compensation on privacy concern and behavioral intention," ACM SIGMIS Database DATABASE Adv. Inf. Syst., vol. 40, no. 1, pp. 38–51, 2009.

[20] J. Phelps, G. Nowak, and E. Ferrell, "Privacy concerns and consumer willingness to provide personal information," J. Public Policy Mark., vol. 19, no. 1, pp. 27–41, 2000.

[21] S. Patil and A. Kobsa, "Uncovering privacy attitudes and practices in instant messaging," in Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work, 2005, pp. 109–112.

[22] T. Dinev and P. Hart, "An extended privacy calculus model for e-commerce transactions," Inf. Syst. Res., vol. 17, no. 1, pp. 61–80, 2006.

[23] S. L. Jarvenpaa, N. Tractinsky, and M. Vitale, "Consumer trust in an Internet store," Inf. Technol. Manag., vol. 1, no. 1–2, pp. 45–71, 2000.

[24] T. Dinev and P. Hart, "Internet privacy concerns and their antecedents-measurement validity and a regression model," Behav. Inf. Technol., vol. 23, no. 6, pp. 413–422, 2004.

[25] H.-M. Kim and G.-W. Bock, "The Role of Attention and Neutralization in Posting Malicious Comments Online," 2018.

[26] K. Shim, "Does Fear of Isolation Disappear Online? Attention-Seeking Motivators in Online Political Engagement," Media Commun., vol. 7, no. 1, pp. 128–138, 2019.

[27] S. T. Hawk, R. J. J. M. van den Eijnden, C. J. van Lissa, and T. F. M. ter Bogt, "Narcissistic adolescents' attention-seeking following social rejection: Links with social media disclosure, problematic social media use, and smartphone stress," Comput. Human Behav., vol. 92, pp. 65–75, 2019.

[28] K.-L. Hui, H. H. Teo, and S.-Y. T. Lee, “The value of privacy assurance: an exploratory field experiment,” Mis Q., pp. 19–33, 2007.

[29] G. R. Milne and M. J. Culnan, “Strategies for reducing online privacy risks: Why consumers read (or don’t read) online privacy notices,” J. Interact. Mark., vol. 18, no. 3, pp. 15–29, 2004.

[30] A. Aljaafreh, A. Al-Ani, R. Aladaileh, and R. Aljaafreh, “Initial trust in internet banking service in Jordan: Modeling and instrument validation,” J. Theor. Appl. Inf. Technol., 2015.

[31] R. K. Chellappa and R. G. Sin, “Personalization versus privacy: An empirical examination of the online consumer’s dilemma,” Inf. Technol. Manag., vol. 6, no. 2–3, pp. 181–202, 2005.

[32] M. J. Culnan and P. K. Armstrong, “Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation,” Organ. Sci., vol. 10, no. 1, pp. 104–115, 1999.

[33] C. Jensen, C. Potts, and C. Jensen, “Privacy practices of Internet users: Self-reports versus observed behavior,” Int. J. Hum. Comput. Stud., vol. 63, no. 1–2, pp. 203–227, 2005.

[34] Hajli, M. N., “A Study of the Impact of Social Media on Consumers”,

[35] International Journal of Market Research, Vol. 56, No. 3, 2014, pp. 387-404.

[36] Hair, B., & Black, J. W. Babin & Anderson, “Multivariate Data Analysis”, 2010.

[37] Cheung, C., Lee, Z. W., & Chan, T. K. (2015). Self-disclosure in social networking sites. Internet Research, 25(2), 279-299.

Investigating Epidemic Growth of COVID-19 in Saudi Arabia based on Time Series Models

Mohamed Torky[1]
Dept. of Computer Science, Higher Institute of Computer Science and Information Systems
Culture and Science City, 6 October City, Giza, Egypt
Scientific Research Group in Egypt (SRGE)

M. Sh Torky[2]
Dept. of General Courses, College of Applied Studies and Community Service, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia
Scientific Research Group in Egypt (SRGE)

Azza. A. A[3]
Computer Science Department
College of Science and Humanitiesin Jubail
Imam Abdulrhaman Bin Faisal University

Aboul Ella Hassanein[4]
Dept. of Information Technology
Faculty of Computers and Artificial Intelligence
Cairo University, Giza, Egypt
Scientific Research Group in Egypt (SRGE)

Wael Said[5]
Dept. of Computer Science
Faculty of Computers and Informatics
Zagazig Univesity, Zagazig, Egypt

Abstract—**Predictive mathematical models for simulating the spread of the COVID-19 pandemic are an interesting and fundamental approach to understand the infection growth curve of the epidemic and to plan effective control strategies. Time series predictive models are one of the most important mathematical models that can be utilized for studying the pandemic growth curve. In this study, three-time series models (Susceptible-Infected-Recovered-Death (SIRD) model, Susceptible-Exposed-Infected-Recovered-Death (SEIRD) model, and Susceptible-Exposed-Infected-Quarantine-Recovered-Death-Insusceptible, (SEIQRDP) model) have been investigated and simulated on a real dataset for investigating Covid-19 outbreak spread in Saudi Arabia. The simulation results and evaluation metrics proved that SIRD and SEIQRDP models provided a minimum difference error between reported data and fitted data. So using SIRD, and SEIQRDP models are used for predicting the pandemic end in Saudi Arabia. The prediction results showed that the Covid-19 growth curve will be stable with detected zero active cases on 2 February 2021 according to the prediction computations of the SEIQRDP model. Also, the prediction results based on the SIRD model showed that the outbreak will be stable with active cases after July 2021.**

Keywords—COVID-19 outbreak; time series models; SIRD; SEIRD; SEIQRDP

I. Introduction

The rapid and continuous spread of the Covid-19 pandemic throughout the world still represents a big dilemma for all countries. An according to the situation report-205 issued by the World Health Organization (WHO) on 12 August 2020, there are more than 20M infected cases of COVID-19 and more than 730,000 deaths globally [1]. King of Saudi Arabia (KSA) is the first largest country in the Arab world that infected with more than 293,000 infected cases and 3,000 deaths due to the COVID-19 Pandemic [2]. While stopping the spread of the infection is becoming an extremely big challenge according to the WHO, countries implemented some strict measures to control the infection growth. In KSA, the government applied some countermeasures such as issuing a social distance app," Tabaud", to notify citizens if they came into contact with an individual infected with COVID-19 [3]. KSA also prevented performing Umrah and sets some COVID-19 protocols, and restrictions for limited Hajj this year. Moreover, the Ministry of Interior issued a package of provisions and penalties for violators of the measures and protocols taken to suppress the pandemic spread.

Although KSA has been considered as of the first countries that took precaution and preventive countermeasures for curbing the COVID-19 outbreak and utilizing all its capabilities to suppress its spread, the taken countermeasures against COVID-19 until this time of writing this paper didn't zeroize the growth of infected cases in the kingdom. The Corona tracker report published on 21 august 2020 informed that KSA has 303,973 confirmed cases and 3548 deaths [4].

Harnessing predictive mathematical models for pandemics [5-8] is necessary and fundamental to trace the epidemic and to plan effective control procedures [9][10]. Predictive mathematical models have long been providing various patterns of quantitative information in outbreaks as well as providing useful recommendations and guidelines to pandemic control and decision making. Hence, investigating epidemiological diseases mathematically becomes a very necessary and important issue [11].

In this study, we try to use three-time series models, Susceptible-Infected-Recovered-Death (SIRD) model, Susceptible-Exposed-Infected-Recovered-Death (SEIRD) model, and the generalized Susceptible-Exposed-Infected-Quarantine-Recovered-Death-Insusceptible (SEIQRDP) model for predicting COVID-19 spread in KSA. The three models have been simulated on a real dataset obtained from [12]. Also, the performance of the three models has been investigated and tested across four periods of time on the used dataset. Then, the three models SIRD, SEIRD, and SEIQRDP have been tested for fitting data of COVID-19 spreading in KSA, and then a selection of the best-fitted models used for predicting the COVID-19 outbreak in KSA. Choosing the best-fitted models is based on calculating the least Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Deviation (MAD) between fitted data and reported data. The simulation and experimental results proved that the SEIQRDP model achieved good prediction results regarding the pandemic growth and end.

The rest of this paper can be organized as: Section 2 discusses the three models, SIRD, SEIRD, and SEIQRDP models. Section 3 presents the simulation and experimental results. Section 4 discusses the obtained results. Finally, Section 5 presents the conclusion and future work.

II. Literature Review

The literature introduced several studies to mathematically study the infection growth of the COVID-19 outbreak. Time series analysis models are common techniques that have been utilized to effectively model, estimate, and predict the growth of the COVID-19 pandemic [13][14]. Zeynep in [15] studied utilizing Auto-Regressive Integrated Moving Average (ARIMA) time series models to the spread of COVID-19 of three European countries most affected by COVID 19: Italy, Spain, and France. This study clarified that ARIMA models are appropriate techniques for forecasting COVID-19 spread in the future and provide a good understanding of the epidemiological stage of these countries.

Elmousalami et al. in [16] investigated three-time series models, moving average (MA), weighted moving average (WMA), and single exponential smoothing (SES) for creating a comparison of day level forecasting models on COVID-19 cases (i.e. confirmed, recovered, and deaths). The three models have been simulated on a real dataset, and the results indicated that the SEIRD model is the most effective and accurate technique for predicting confirmed, recovered, and death cases COVID-19 in Egypt.

Cooper et al. in [17] studied to study the effectiveness of modeling COVID-19 spread in different countries using the Susceptible-Infected-Removed (SIR) model. The simulation results showed the importance of modeling COVID-19 spread by the SIR model that can assist to estimate the impact of the pandemic by offering valuable predictions results in China, South Korea, India, Australia, USA, Italy, and the state of Texas in the USA.

Also, the SIR model can be used for investigating the predicting the peak and the end of the epidemic, the effect of the asymptomatic infection on the spread of COVID-19 outbreak, herd immunity variables, and social distance parameters [18].

The extended version of the SIR model is the Susceptible-Exposed-Infectious-Recovered-Dead (SEIRD) model that can be used also as another time series model for modeling COVID-19 spread [19]. Maguire et al in [20] used the SEIRD model for modeling COVID-19 spread in Sicily, Italy. The experimental results showed a good fit between reported data and estimated data using the SEIRD model.

The fractional-order differential equations add extra solutions in the study of the COVID-19 outbreak. So, the fractional version of many epidemical models have been studied in various works as in [5] and [21-23].

III. Time Series Prediction Models

This section discusses three selected time series prediction models we used for predicting the epidemic growth of COVID-19 in Saudi Arabia.

In this paper, the generalized SEIQRDP model is used to visualize the epidemic growth of COVID-19 in Saudi Arabia with a comparative analysis with two models SEIRD and SIRD. SEIRD and SIRD models have been derived from the generalized SEIQRDP model. The following subsections give more explanation of the three models.

A. *Susceptible, Exposed, Infection, Quarantined, Recovery, Death, Insusceptible (SEIQRDP) Model*

In the SEIQRDP model (Fig. 1), seven different states of infection transition can be considered in different analysis manner:

- Susceptible cases S(t).
- Insusceptible cases P(t).
- Exposed cases (t).
- Infectious cases (t).
- Quarantined cases (t).
- Recovered cases R(t).
- Dead cases (t).

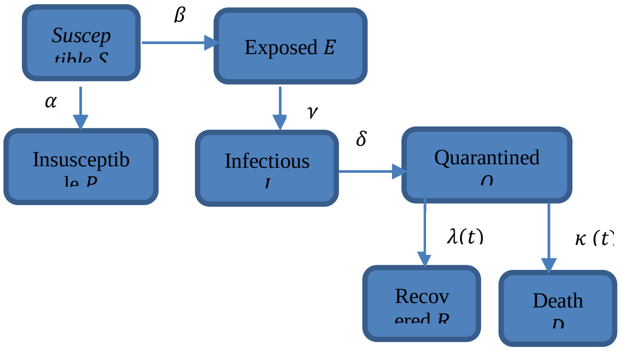


Fig. 1. The Generalized SEIQRDP Model.

Also, it depends on six parameters:

- α: Protection rate.
- β: Infection rate.
- γ^{-1}: Inverse of the average latent time.
- δ^{-1}: Quarantine rate at which infectious people enter in quarantine.
- **λ(t)**: Recovery rate, time-dependent recovery rate.
- κ(t): Mortality rate, time-dependent mortality rate.

This model depends on six infection transition equations that can be used for studying the infection spread of COVID-19, Susceptible cases (t), Infectious cases $I(t)$, Exposed cases $E(t)$, Quarantined cases $Q(t)$, Recovered cases $R(t)$, and Dead cases $D(t)$. SEIQRDP system equations are given in (1).

$$\frac{ds}{dt} = -\beta \frac{S(t)I(t)}{N_{pop}} - \ \alpha\, s(t)$$

$$\frac{dE}{dt} = \beta \frac{S(t)I(t)}{N_{pop}} - \gamma E(t)$$

$$\frac{dI}{dt} = \gamma E(t) - \delta I(t)$$

$$\frac{dQ}{dt} = \delta I(t) - \lambda(t)Q(t) - \kappa(t)Q(t) \tag{1}$$

$$\frac{dR}{dt} = \lambda(t)I(t)$$

$$\frac{dD}{dt} = \kappa\ (t)I(t)$$

$$\frac{dP}{dt} = \alpha\, s(t)$$

B. *Susceptible, Infection, Recovery, Death (SIRD) Model*

This model depends on four infection transition equations that can be considered in the mathematical analysis of studying COVID-19 spread, Susceptible cases(t), Infectious cases $I(t)$, Recovered cases $R(t)$, and Dead cases $D(t)$(Fig. 2). We modified the generalized SEIQRDP system equations model to produce the SIRD system equation model in (2).

$$\frac{ds}{dt} = -\beta \frac{S(t)I(t)}{N_{pop}}$$

$$\frac{dI}{dt} = \beta \frac{S(t)\ I(t)}{N_{pop}} - \lambda(t)\ I(t) - \ k(t)\ I(t) \tag{2}$$

$$\frac{dR}{dt} = \lambda(t)\ I(t)$$

$$\frac{dR}{dt} = \lambda(t)I(t)$$

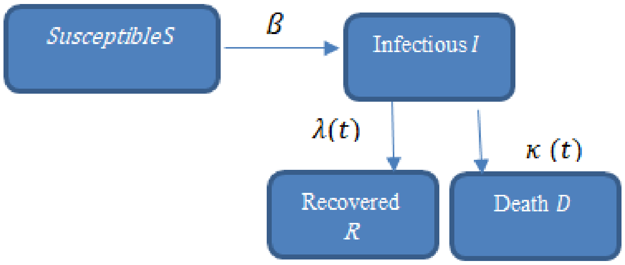


Fig. 2. The SIRD Model.

C. *Susceptible, Exposed, Infection, Recovery, Death (SEIRD) Model*

This model depends on five infection transition equations that can be used for studying the infection spread of COVID-19, Susceptible cases (t), Infectious cases $I(t)$, Exposed cases $E(t)$, Recovered cases $R(t)$, and Death cases $D(t)$ (Fig. 3).

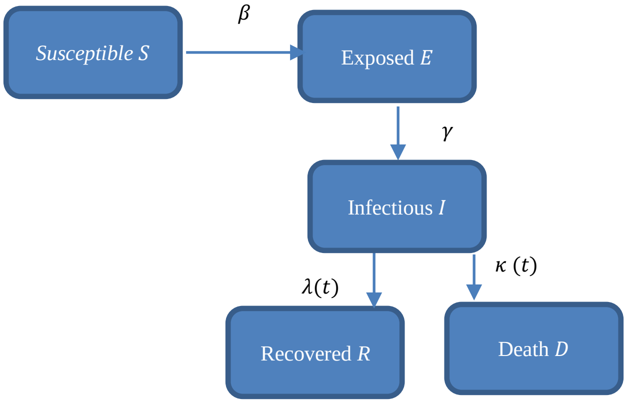


Fig. 3. The SEIRD Model.

We modified the generalized SEIQRDP system equations model to produce the SEIRD system equation model in Eq.3.

$$\frac{ds}{dt} = -\beta \frac{S(t)I(t)}{N_{pop}}$$

$$\frac{dE}{dt} = \beta \frac{S(t)I(t)}{N_{pop}} - \gamma E(t)$$

$$\frac{dI}{dt} = \gamma E(t) - \lambda I(t) - \kappa\ I(t)\ (3)$$

$$\frac{dR}{dt} = \lambda(t)I(t)$$

$$\frac{dD}{dt} = \kappa(t)\ I(t)$$

IV. Simulation and Experimental Results

The three mathematical models, SIRD, SEIRD, and SEIQRDP have been applied to the Saudi Arabia data set collected from [12]. The data set presents several active cases (i.e. infected cases and still infected), recovered cases, and death cases in Saudi Arabia between 2 February 2020 and 10 August 2020. Fig. 4 depicts and visualizes the three classes of our data set. The three models have been tested using three metrics, Mean Absolute deviation (MAD), Mean Square Error

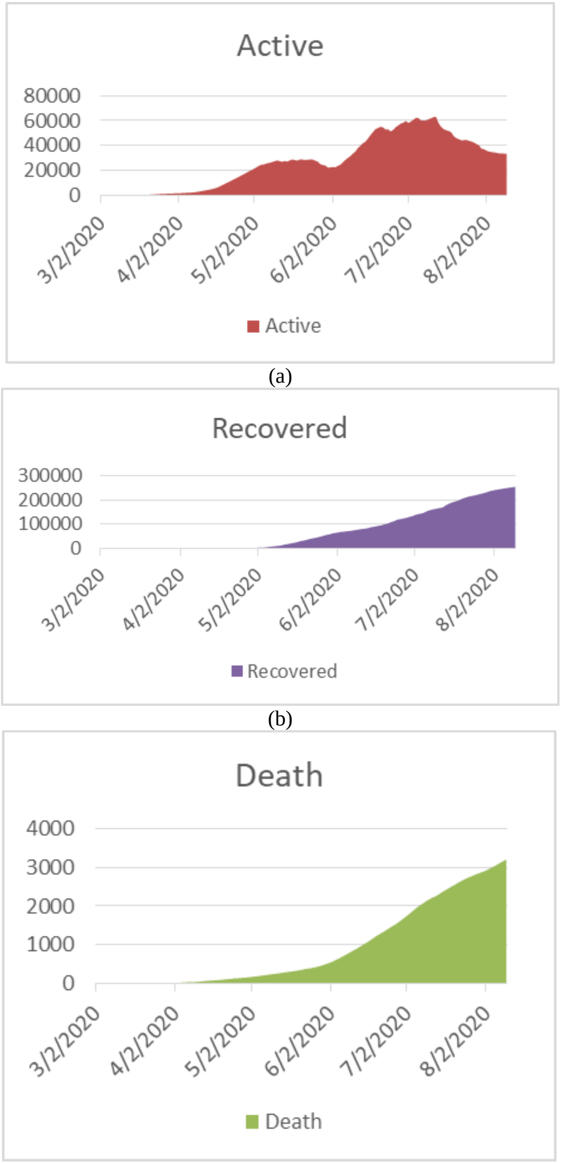
(MSE), and Mean Absolute Percentage Error (MAPE), (as depicted in equations 4, 5, and 6 to investigate their fitness against the reported data. The investigation has been done within four periods, (from 24/3/2020 (where the death cases are reported) to 10/8/2020), (from 1/4/2020 to 10/8/2020), (from 1/5/2020 to 10/8/2020), and (from 15/6/2020 to 10/8/2020.

$$MAD = \frac{1}{n}\sum_{i=1}^{n}[Y_i - \tilde{Y}_i]^1 \quad (4)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}[Y_i - \tilde{Y}_i]^2 \quad (5)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{[Y_i - \tilde{Y}_i]}{\tilde{Y}_i} \times 100 \quad (6)$$

Where n is the total case in the data class (i.e. infected, recovered, death), i is the number of the case and $\tilde{Y}_i$ is the predicted outcome of the time series model.



(a)

(b)

(c)

Fig. 4. Dataset Visualization: (a) Active Cases, (b) Recovered Cases, and (c) Death Cases.

A. Testing SIRD, SEIRD, and SEIQRD Models from 24/3/2020 to 10/8/2020

The actual reported data of recovered, active, and death cases (from 24/3/2020 to 10/8/2020) has been compared with the fitted recovered, active, and death cases results while applying the three models, SIRD, SEIRD, and SEIQRDP. Table I summarizes and compares the testing results using Mean Absolute deviation (MAD), Mean Square Error (MSE), and Mean Absolute Percentage Error (MAPE). Also, Fig. 5 visualizes and compares the fitted results of the three models from 24/3/2020 to 10/8/2020.

B. Testing SIRD, SEIRD, and SEIQRDP Models from 1/4/2020 to 10/8/2020

The actual reported data of recovered, active, and death cases (from 1/4/2020 to 10/8/2020) has been compared with the fitted recovered, active, and death cases results of the three models, SIRD, SEIRD, and SEIQRDP. Table II summarizes and compares the testing results, and Fig. 6 visualizes and compares the fitted results of the three models from 1/4/2020 to 10/8/2020.

C. Testing SIRD, SEIRD, and SEIQRDP Models from 1/5/2020 to 10/8/2020

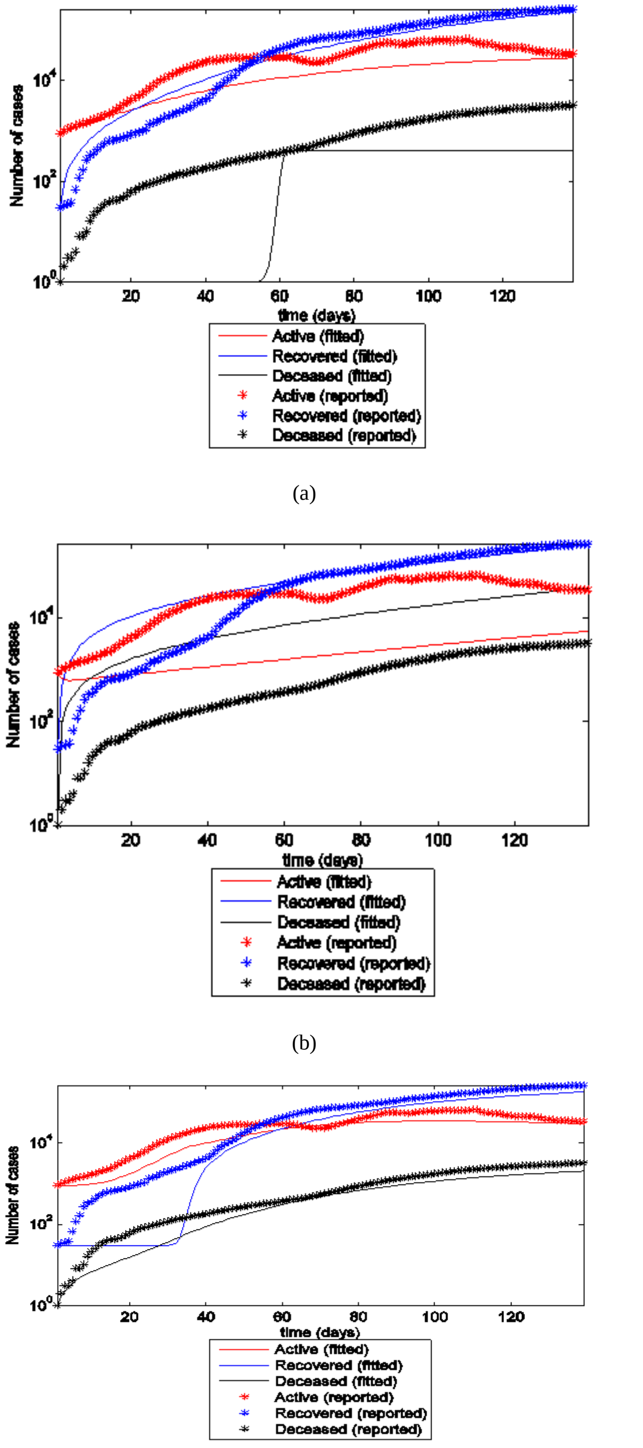
The actual reported data of recovered, active, and death cases (from 1/5/2020 to 10/8/2020) has been compared with the fitted recovered, active, and death cases results of the three models, SIRD, SEIRD, and SEIQRDP. Table III summarizes and compares the testing results, and Fig. 7 visualizes and compares the fitted results of the three models from 1/5/2020 to 10/8/2020.

D. Testing SIRD, SEIRD, and SEIQRDP Models from 15/6/2020 to 10/8/2020

The actual reported data of recovered, active, and death cases (from 15/6/2020 to 10/8/2020) has been compared with the fitted recovered, active, and death cases results of the three models, SIRD, SEIRD, and SEIQRDP. Table IV summarizes and compares the testing results, and Fig. 8 visualizes and compares the fitted results of the three models from 15/6/2020 to 10/8/2020.

TABLE I. TESTING SIRD, SEIRD, AND SEIQRDP MODELS WITHIN THE PERIOD FROM 24/3/2020 TO 10/8/2020

Testing Metric	Data Classes	SIRD	SEIRD	SEIQRDP
MAD	Active cases	1.7620e+004	2.8462e+004	1.4582e+005
	Recovered cases	1.6575e+004	5.2565e+003	2.7677e+004
	Death cases	812.3347	1.5889e+004	355.5890
MSE	Active cases	4.6931e+008	1.1394e+009	2.8040e+010
	Recovered cases	5.6133e+008	3.9453e+008	1.4583e+009
	Death cases	1.4754e+006	4.2800e+008	2.7842e+005
MAPE	Active cases	137.5120	1.2585e+003	85.6326
	Recovered cases	40.4738	40.9124	737.6416
	Death cases	N/A	N/A	N/A



(a)

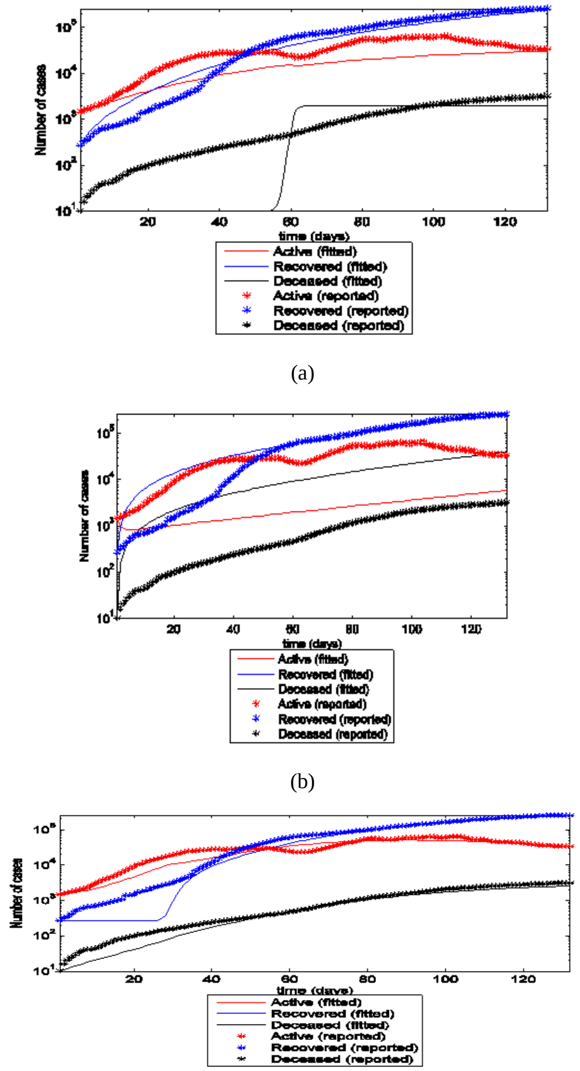
(b)

(c)

Fig. 5. Fitness Testing Results (Active, Recovered, Death) within the Period from 24/3/2020 to 10/8/2020. (a) SIRD Model, (b) SEIRD Model, and (c) SEIQRDP Model.

TABLE II. TESTING SIRD, SEIRD, AND SEIQRDP MODELS WITHIN THE PERIOD FROM 1/4/2020 TO 10/8/2020

Testing Metric	Data Classes	SIRD	SEIRD	SEIQRDP
MAD	Active cases	1.6375e+004	2.9792e+004	2.8625e+004
	Recovered cases	1.4029e+004	- 2.4986e+003	5.8415e+003
	Death cases	18.6632	- 1.9001e+004	162.9914
MSE	Active cases	4.0328e+008	1.1829e+009	1.0949e+009
	Recovered cases	3.8735e+008	1.5614e+008	6.9528e+007
	Death cases	3.9091e+005	5.9562e+008	6.2496e+004
MAPE	Active cases	106.5099	1.1252e+003	862.8085
	Recovered cases	31.8298	32.2157	88.8226
	Death cases	699.9287	94.7098	39.6062



(a)

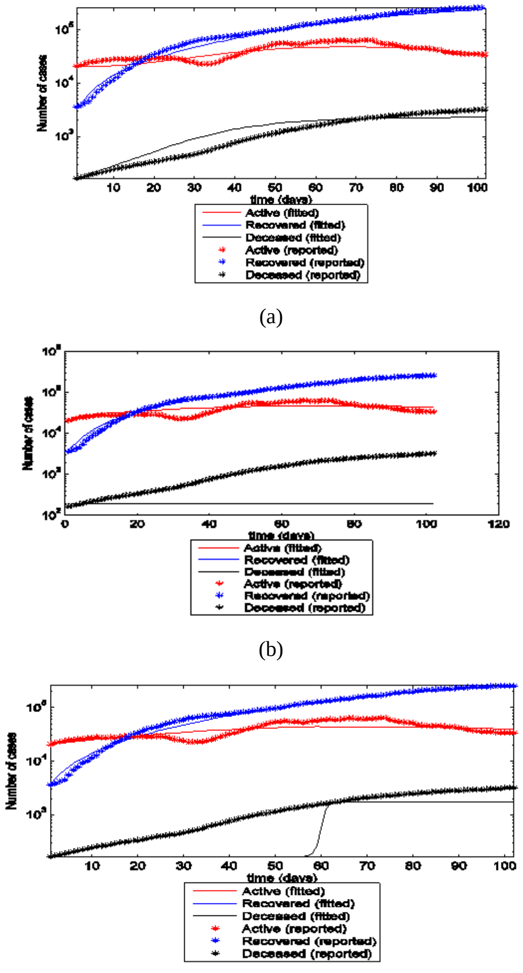
(b)

(c)

Fig. 6. Fitness Testing Results (Active, Recovered, Death) within the Period from 1/4/2020 to 10/8/2020. (a) SIRD Model, (b) SEIRD Model, and (c) SEIQRDP Model.

TABLE III. TESTING SIRD, SEIRD, AND SEIQRDP MODELS WITHIN THE PERIOD FROM 1/5/2020 TO 10/8/2020

Testing Metric	Data Classes	SIRD	SEIRD	SEIQRDP
MAD	Active cases	2.3503e+003	-1.0001e+003	1.7735e+004
	Recovered cases	3.9296e+003	-2.3425e+004	9.0101e+003
	Death cases	592.7127	1.2233e+003	-100.6245
MSE	Active cases	8.3956e+007	2.1594e+008	4.2607e+008
	Recovered cases	3.6489e+007	8.1871e+008	1.2748e+008
	Death cases	5.4580e+005	2.4872e+006	1.9323e+005
MAPE	Active cases	18.0708	24.2915	92.4594
	Recovered cases	8.4338	21.7347	11.5785
	Death cases	179.1255	755.1259	26.1291



(a)

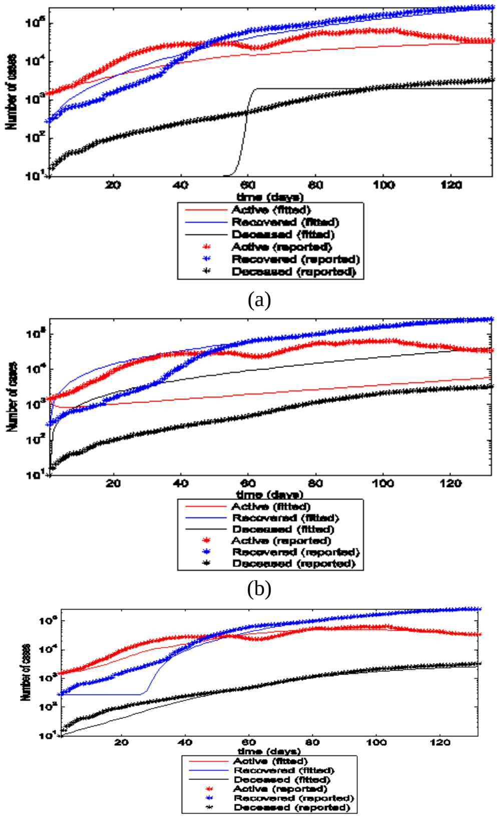
(b)

(c)

Fig. 7. Fitness Testing Results (Active, Recovered, Death) within the Period from 1/5/2020 to 10/8/2020. (a) SIRD Model, (b) SEIRD Model, and (c) SEIQRDP Model.

TABLE IV. TESTING SIRD, SEIRD, AND SEIQRDP MODELS WITHIN THE PERIOD FROM 15/6/2020 TO 10/8/2020

Testing Metric	Data Classes	SIRD	SEIRD	SEIQRDP
MAD	Active cases	3.4232e+003	4.5401e+004	4.6662e+004
	Recovered cases	1.9706e+003	-2.6146e+004	-1.2491e+003
	Death cases	-718.9346	-2.0110e+004	-0.0685
MSE	Active cases	2.4548e+007	2.1908e+009	2.2523e+009
	Recovered cases	1.3615e+007	7.3977e+008	7.1920e+006
	Death cases	3.7679e+006	4.6736e+008	985.3657
MAPE	Active cases	8.0891	1.4721e+003	1.7721e+003
	Recovered cases	1.9733	14.2931	1.0960
	Death cases	56.6488	88.5107	1.2266



(a)

(b)

(c)

Fig. 8. Fitness Testing Results (Active, Recovered, Death) within the Period from 15/6/2020 to 10/8/2020. (a) SIRD Model, (b) SEIRD Model, and (c) SEIQRDP Model.

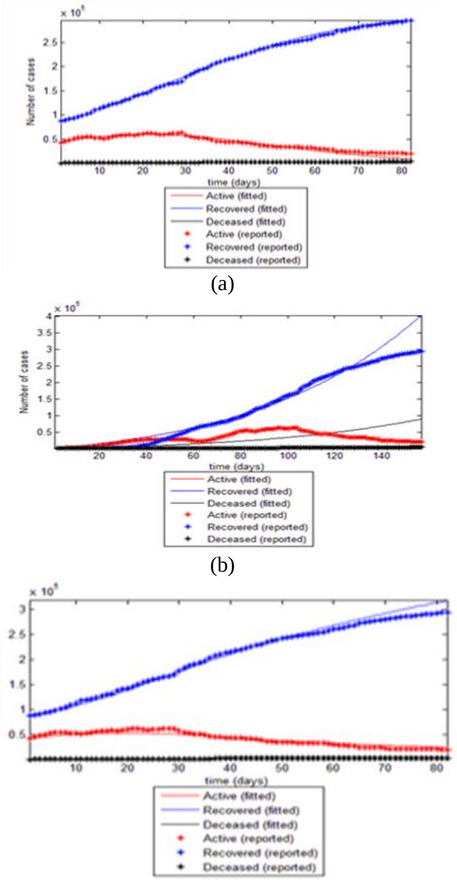
V. PREDICTING COVID-19 OUTBREAK END IN SAUDI ARABIA

In this section, the three models, SIRD, SEIRD, and SEIQRDP have been simulated again to evaluate their prediction computations compared to the observed values within the period from 11/8/2020 to 4/9/2020. The prediction computation has been based on the best period within which each model provided the least Mean Square Error (MSE) of fitting data as summarized in Table V. Fig. 9 depicts the prediction results of the three models within the period 11/8/2020 to 4/9/2020, a) SIRD model, b) SEIRD model, and c) SEIQRDP model.

According to these results, the SIRD and SEIQRDP are the best models that provided good prediction results within the period 11/8/2020 to 4/9/2020 compared to the observed data.

TABLE V. THE AVERAGE OF THE LEAST MEAN SQUARE ERROR (MSE) OF THE THREE MODELS BASED ON THE BEST PERIOD OF FITTING DATA FOR EACH MODEL WHEN APPLIED WITHIN THE PERIOD OF PREDICTION FROM 11/8/2020 TO 4/9/2020

	SIRD	SEIRD	SEIQRDP
Best Period of fitting data	15/6/2020 to 10/8/2020	1/4/2020 to 10/8/2020	1/4/2020 to 10/8/2020
Period of Prediction	11/8/2020 to 4/9/2020		
MSE for Active cases	2.3849e+007	1.0393e+009	1.7434e+009
MSE for Recovered cases	7.1894e+007	7.6838e+008	9.4376e+006
MSE for Death cases	4.6849e+006	1.3485e+009	2.2081e+004
Average of MSE	3.35E+07	1.05E+09	5.84E+08



(a)

(b)

(c)

Fig. 9. The Prediction Results of the Three Models within the Period 11/8/2020 to 4/9/2020, (a) SIRD Model, (b) SEIRD Model, and (c) SIEQRDP Model.

Hence, we used these two models to predict the end of the COVID-19 outbreak in Saudi Arabia. According to the prediction results of the SIRD model, the SIRD curve begins stable regarding detecting new active cases of Covid-19 after July 2021 as depicted in Fig. 10. Also, according to the prediction results of the SIEQRDP model, the SIEQRD curve begins stable with Zero active cases of Covid-19 at 2/2/2020 as depicted in Fig. 11.

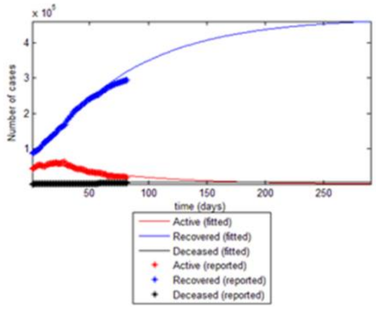


Fig. 10. SIRD Curve begin Stable with Active cases after July 2021.

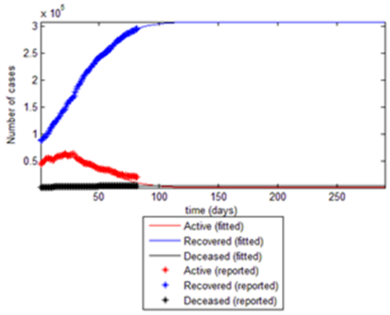


Fig. 11. SEIQRDP Curve begin Stable with Zero Active Cases at 2/2/2021.

VI. DISCUSSION

Prior studies that have noted the importance of using various mathematical models for predicting the spread of COVID-19 pandemic are different. Most of these studies have been established based on time series models [13-20], and other studies have been based on differential equations models [5] [21-23]. However, very little effort was found in the literature on the question of predicting the end of the COVID-19 outbreak in Saudi Arabia.

The most interesting finding was that the SIRD and SIEQRD are the best models for predicting the pandemic growth and end in Saudi Arabia based on evaluating the average of Mean Square Error (MSE). The results clarified that the SIEQRD curve begins stable with zero active cases at 2/2/2021as depicted previously in Fig. 11. Another important finding was that the SIRD curve begin stable with active cases of Covid-19 after July 2021 as depicted previously in Fig. 10.

However, the SEIRD model doesn't provide satisfactory prediction results compared to SIRD and SEIQRDP models.

These results seem to be consistent with other research that found that the SIRD model [24][25] and SEIQRDP model [26] are effective models for predicting Covid-19 pandemic growth. However, this study is contrary to [19] [20] that defends upon the SEIRD model as a good prediction model for estimating the pandemic growth in Italy. This inconsistency may be due to the difference in the features of the data set while the model is simulating on different datasets. Hence, these findings may be somewhat limited by the features of the datasets and simulation environment. The present results are significant in at least two major respects, comparing three prediction models for predicting the pandemic growth curve of the COVID-19 outbreak in Saudi Arabia, and predicting the end of the pandemic based on the best prediction model.

A further study with more focus on using differential equation-based models is therefore suggested to study the Covid-19 outbreak growth and predicting its end by conducting different simulations with different datasets in different countries.

VII. Conclusion

The present study was designed to investigate applying three-time series models for studying COVID-19 growth in Saudi Arabia. The study simulated the mathematical systems of SIRD, SEIRD, and SEIQRDP on a real dataset of Saudi Arabia. The study presented a set of comparative analyses on the used dataset for investigating and evaluating the effectiveness of the three models in predicting the COVID-19 pandemic growth as well as predicting the end date of this outbreak. The finding of this study clarifies that SIRD and SEIQRDP models provide good prediction results about the pandemic growth and its end date in Saudi Arabia. The prediction results showed that the Covid-19 growth curve will be stable with detected zero active cases on 2 February 2021 according to the prediction computations of the SEIQRDP model. Also, the prediction results based on the SIRD model showed that the outbreak will be stable with the detected active cases after July 2021.

This new understanding should help to improve predictions of the impact of using SIRD and SIEQRD models for studying the COVID-19 growth curve in different datasets that have various infection dynamics in different countries.

For more prediction accuracy, a further study with more focus on using differential equation-based models is needed to study the Covid-19 outbreak growth and predicting its end. This can be achieved by conducting different simulations on different datasets in different countries using some differential equations-based models.

References

[1] WHO, Coronavirus disease (COVID-19) Situation Report – 205 , [online], Available at: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200812-covid-19-sitrep-205.pdf?sfvrsn=627c9aa8_2 (Access 12/8/2020).

[2] Worldometers, Saudi Arabia, [Online], Available at https://www.worldometers.info/coronavirus/country/saudi-arabia/ (access 12/8/2020).

[3] Tuka K, Saudi Arabia’s coronavirus social distancing app ‘Tabaud’: All you need to know, (28 June 2020), [online], Available at: https://english.alarabiya.net/en/coronavirus/2020/06/28/Saudi-Arabia-s-coronavirus-social-distancing-app-Tabaud-All-you-need-to-know (access 21/8/2020).

[4] Corona Tracker, Saidi Arabia Overview, [online], Available at https://www.coronatracker.com/country/saudi-arabia/ (access 21/8/2020).

[5] Higazy M. Novel Fractional Order SIDARTHE Mathematical Model of the COVID-19 Pandemic. Chaos, Solitons & Fractals. Vol 138, 2020 Jun 13:110007.

[6] Ball FG, Knock ESPD. O’Neil Control of emerging infectious diseases using responsive imperfect vaccination and isolation. Math. Biosci. 20 08;216(1):10 0–13.

[7] Laarabi H, Abta A, Hattaf K. Optimal Control of a delayed SIRS epidemic model with vaccination and treatment. Acta Biotheor 2015;63(15):87–97.

[8] Shim E, Tariq A, Choi W, Lee Y, Chowell G. Transmission potential and severity of COVID-19 in South Korea. International Journal of Infectious Diseases. 93 (2020), 339-344.

[9] D'Arienzo M, Coniglio A. Assessment of the SARS-CoV-2 basic reproduction number, R0, based on the early phase of COVID-19 outbreak in Italy. Biosafety and Health. 2020 Apr 2.

[10] Liang K. Mathematical model of infection kinetics and its analysis for COVID-19, SARS, and MERS. Infection, Genetics, and Evolution. 82 (2020) 104306.

[11] Jewell NP, Leonard JA, Jewell BL. Predictive mathematical models of the COVID-19 pandemic: Underlying principles and value of projections. Jama. 2020 May 19;323(19):1893-4.

[12] CSSEGISandData/COVID-19, [online], https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series (access 12/6/2020).

[13] Maleki M, Mahmoudi MR, Heydari MH, Pho KH. Modeling and forecasting the spread and death rate of coronavirus (COVID-19) in the world using time series models. Chaos, Solitons & Fractals. 2020 Jul 25:110151.

[14] Maleki M, Mahmoudi MR, Wraith D, Pho KH. Time series modeling to forecast the confirmed and recovered cases of COVID-19. Travel Medicine and Infectious Disease. 2020 May 13:101742.

[15] Ceylan Z. Estimation of COVID-19 prevalence in Italy, Spain, and France. Science of The Total Environment. 2020 Apr 22:138817.

[16] Elmousalami HH, Hassanien AE. Day level forecasting for Coronavirus Disease (COVID-19) spread: analysis, modeling, and recommendations. arXiv preprint arXiv:2003.07778. 2020 Mar 15.

[17] Cooper I, Mondal A, Antonopoulos CG. A SIR model assumption for the spread of COVID-19 in different communities. Chaos, Solitons & Fractals. 2020 Jun 28:110057.

[18] Chen YC, Lu PE, Chang CS, Liu TH. A Time-dependent SIR model for COVID-19 with undetectable infected persons. arXiv preprint arXiv:2003.00122. 2020 Feb 28.

[19] Fonseca i Casas P, García i Carrasco V, Garcia i Subirana J. SEIRD COVID-19 Formal Characterization and Model Comparison Validation. Applied Sciences. 2020 Jan;10(15):5162.

[20] Maugeri A, Barchetta M, Battiato S, Agodi A. Modeling the Novel Coronavirus (SARS-CoV-2) Outbreak in Sicily, Ital., Int. J. Environ. Res. Public Health 2020, 17(14), 4964.

[21] Kulkarni S, Takale K, Shaikh A. Application of Adomian decomposition method to solve the fractional mathematical model of coronavirus. J. Math. Comput. Sci.. 2020 May 26;10(5):1327-39.

[22] Khan MA, Atangana A. Modeling the dynamics of novel coronavirus (2019- nCov) with fractional derivative. Alexandria Engineering Journal 2020:1–11. https://doi.org/10.1016/j.aej.2020.02.033.

[23] Baleanu D, Mohammadi H, Rezapour S. A fractional differential equation model for the COVID-19 transmission by using the Caputo–Fabrizio derivative. Advances in difference equations. 2020 Dec;2020(1):1-27.

[24] Fernández-Villaverde J, Jones CI. Estimating and Simulating a SIRD Model of COVID-19 for Many Countries, States, and Cities. National Bureau of Economic Research; 2020 May 7.

[25] Lalwani S, Sahni G, Mewara B, Kumar R. Predicting optimal lockdown period with a parametric approach using three-phase maturation SIRD model for COVID-19 pandemic. Chaos, Solitons & Fractals. 2020 May 30:109939.

[26] Caccavo D. Chinese and Italian COVID-19 outbreaks can be correctly described by a modified SIRD model. medRxiv. 2020 Jan 1.

Learners' Activity Indicators Prediction in e-Learning using Fuzzy Logic

Sanae CHEHBI[1], Chakir FRI[3]
PHD Student at Computer Science Department Faculty of Sciences, Moulay Ismail University of Meknes
Meknes, Morocco

Rachid ELOUAHBI[2]
Higher Education Professor Department of Computer Science, Faculty of Sciences, Moulay Ismail University of Meknes, Meknes, Morocco

Abstract—**With the idea of introducing computer supports in education, Online Learning (named also e-learning) associated on one hand, the concept of network, therefore that of distance and concepts of communicating interaction, whether between the learner and the teacher (or tutor), or between the learners themselves and on the other hand exchanges and collaboration. Any activity in e-learning leaves recorded traces stored in a database system. Until now, data on student activity is stored as low-level information; however, the volume of this information is too large to be processed and interpreted by tutors, requiring data collection and preparation to give it meaning. In addition, according to the studies carried out in this direction, the tracking of learners must be guaranteed in all stages of e-learning process, to assist and help them when they encounter problems that they cannot solve. The lack of direct contact between the tutor and the learners can cause a lack of feedback of the learning activity; all these problems can lead to a high rate of abundance in e-learning. Our work aims to develop a model for predicting learner activity indicators using fuzzy logic without going through rigid calculations but based on consultation traces and skill assessment scores. Based on the traces collected from the Learning Management System (LMS) Moodle, it could give the tutor high level processing of the learning activity.**

Keywords—e-Learning; tracking; Moodle; traces; activity indicators; fuzzy logic

I. Introduction

To give learners more choices of methods and means of learning, e-learning offers several advantages such as: accessibility, personalization, adaptability and profitability which qualify it as the best solution. Diversified forms of learning make it possible to increase the level of knowledge acquisition.

During the learning activity, learners leave data from which one can deduce interesting facts and describe well the learning process, called learning traces. These traces are recorded and stored in the LMS database system as low level information [1]. This information is abundant on interactions between students and teachers as well as on access to resources and the system. They can give an idea of how and when students perform their homework and tasks, their engagement in lessons, etc. In a context of e-learning and based on learning traces, teachers have a very partial view of the activity of learners and can on this basis, make judgments about the activity of the learner within the learning system [2].

However, the large volume of this information requires the collection of these traces, preparation and processing for this provided information to be meaningful and transformed into actionable knowledge, which is a difficult task [3]. As Peraya (2019) underlines, learning traces have significant potential for use, whether it is to predict certain learner behaviors, to visualize learning activities or to set up adaptive systems [4]. Authors have got the idea to develop a fuzzy logic system for learners' activity indicators prediction, to assist tutor in tracking learners and give him a clear idea about the course of the learning activity and help him in decision making.

The paper is organized by following three sections. The first sections is to start by the collection and the transformation of traces from the Learning Management System (LMS) Moodle. The second one is to introduce fuzzy logic and its use and contribution in the context of activities' indicators in e-learning LMS. The final sections is to presenting the implementation and simulation of the fuzzy activity indicators system.

II. Tracking Learners in e-Learning

A. e-Learning

e-Learning offers several advantages, among which we cite: ease of access to information, flexibility and autonomy in time and space, variety of learning methods, personalization and individualization of content.

But, it has some limitations such as: the absence of human contact between learners and teacher, the need for greater passion and more rigorous work discipline.

B. Tracking Learners

Capturing the attention of learners is a major challenge whether in "classroom" or in "e-learning". By using an LMS, the teacher does not interact directly with the learners to determine whether they have understood the course well or whether the course is suitable for their levels. This is the reason why researchers introduced tracking in LMS [5].

The crucial task of tutors is to reflect the evolution of the learners, to measure the quality of the training, to highlight the specific knowledge built by each one. This is why it is necessary to collect a set of information on the educational pathways of the learners.

Any activity carried out through software leaves traces recorded in a file, from which one can deduce interesting facts. Since training in Information Technology (IT) has existed, researchers have collected and analyzed the traces left by learners during an e-learning session.

Romero wondered if traces always used for the benefit of the learner? What happens when the traces can be used to justify a negative assessment of the learner's engagement or to assess their participation in a team activity? [3].

1) Trace definition: Among all definitions given to what is a trace, we opt for "The trace carries meaning as any structured representation of inscriptions of knowledge can be (a classification, a reading) with the only difference that this structuring is not voluntary, declarative, but induced by the use of the inscriptions itself; as such, the trace is itself considered as an inscription of knowledge" [6].

C. Background and Related Works

Up to now, a large number of specific learners' tracking tools have been developed to solve online educational problems. These tools are used in different educational environments: Learning and Management Systems (LMS), Intelligent Tutoring Systems (ITS), Adaptive and Intelligent Hypermedia Systems (AIHS), MOOC (Massive Open Online Course). Thus, it is impossible to review or compare our approach against others that already exist. To learn more, you will find a brief description of the specific tools for Moodle and others below.

Florian et al., Focused on analyzing social perspectives to access Moodle tracking data. They tried to reuse the tracking data from the Moodle LMS so that they could model learners and groups. To create rich learning models they used data from Moodle diaries to gain insight into learner activities in a social context and to ensure support for learning. The authors implemented architecture to have a flexible and extensible interface to Moodle's tracking data to transform the collected data into learning analytical information [7].

Conde and al., conducted a study that describes the different existing tools that facilitate the extraction and analysis of educational data for learning analysis. It focuses on two differentiated parts namely: a presentation of the different learning analysis tools that analyzed it, then a work that covers the results, including a comparison of the tools after analyzing the data sets of existing courses, this study summarized the results of the application of the different tools [3].

Iglesias-Pradas, Ruiz-de-Azcárate and Agudo-Peregrina explored the applicability of learning analysis for predicting the development of two transversal skills: teamwork and engagement [8]. They were based on the analysis of logs of Moodle interaction data for the benefit of master's students. The results of the study call into question the appropriateness of the approach and show no relationship between online activity metrics and teamwork and engagement acquisition.

Djouad and Mille proposed to use the traces of the learning activity to develop an indicator management system based on traces (TB-IMS) independent of the platform. The approach adopted made it possible to create and reuse learning indicators independently of the source code of the learning platform. This work presents the theory and its implementation in a first TB-IMS which is illustrated through a real learning situation in Moodle [9].

Gamie and al., introduced a model to analyze and predict students' performance based on two dimensions; teaching style, and e-learning activities. They collected data from educational settings within an academic institution. They analyzed data used to reveal knowledge and useful patterns from which critical decisions could be made [10].

Hernandez-Garcia et al., Proposed the design of a data extraction, transformation and loading (ETL) system of educational data from the LMS Moodle. Their design was based on indicators of teamwork at the individual and team level across four dimensions: communication, cooperation, coordination and monitoring / tracking. This design aims to transform data from teamwork activities, retrieved from Moodle, into useful information about teamwork behaviors [11].

Mazza et al., developed the GISMO tool to extract monitoring data from an online course, it queries the learning traces recorded by the Moodle platform and automatically generates various graphic representations that can be used by tutors. In order to provide visualizations of the course's behavioral, cognitive and social data, allowing constant monitoring of student activities, engagement and learning outcomes [12].

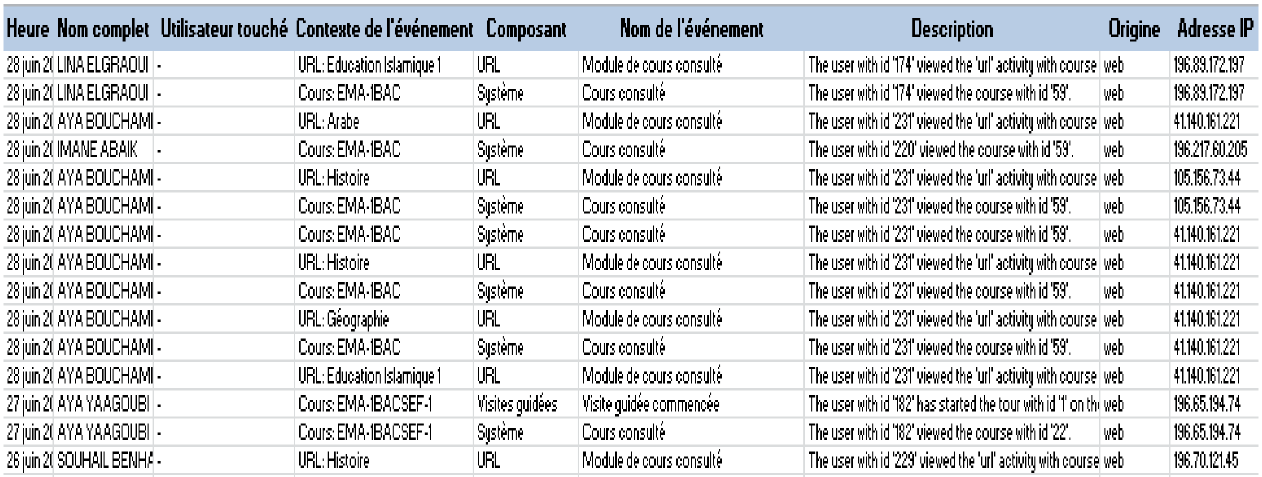
From what precedes, it is obvious that, if there are differences between all these researches in the axes treated and the methods of resolution proposed, they have as a common objective the improvement of learning in order to encourage tutors to make decisions geared towards improving the learning process.

D. Moodle Traces Collection

The during an e-learning session using an LMS, every user is supposed to leave traces of his activity recorded in specific files, from which it is sometimes possible to deduce some interesting facts. This is why the researchers worked on the collection, preparation, processing and analysis of the traces left during the use of the LMS, to ensure the follow-up of learners and to improve the systems put in place. They draw attention to the wealth of indicators extracted which possibly return relevant information on the progress of the learning activity.

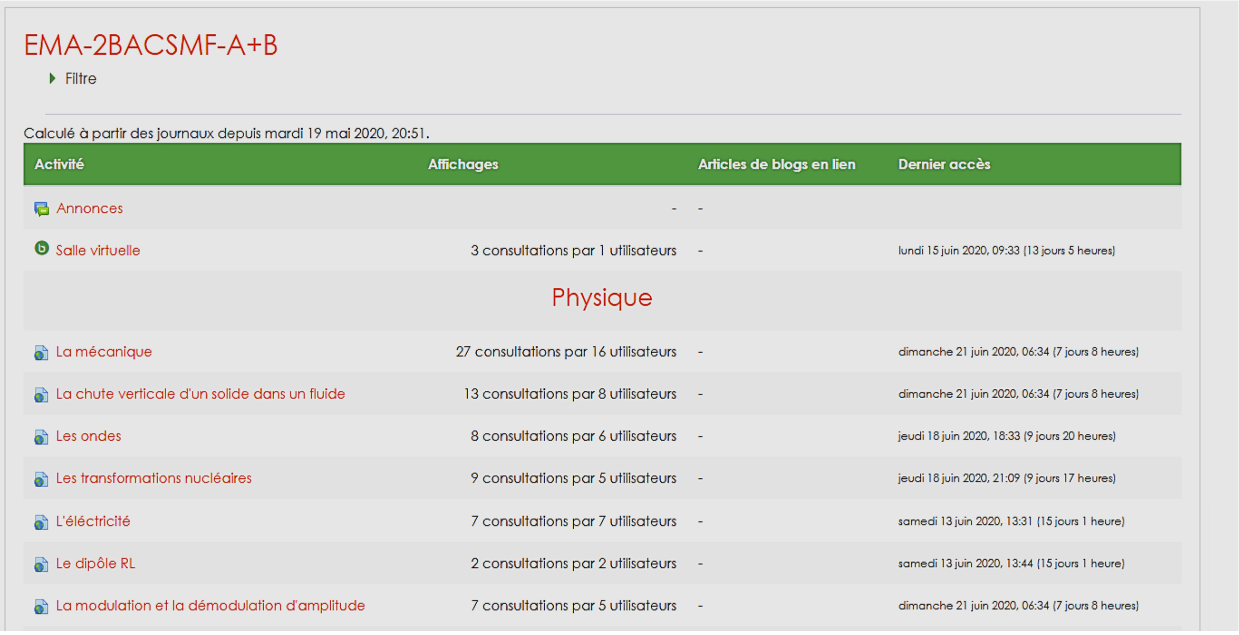
To collect traces we have conducted a study during confinement in a secondary school with students in their final year. In the Fig. 1, the connection log files from the LMS. It contains the session information such as: connection time, the event context and name, the component, the description, the origin of the connection and the IP address.

When consulting the participation of learners in the course, we get a table that contains the number of consultations by the number of learners of a given resource and the last access to the resource, as shown in Fig. 2.



Heure	Nom complet	Utilisateur touché	Contexte de l'événement	Composant	Nom de l'événement	Description	Origine	Adresse IP
28 juin 20	LINA ELGRAOUI	-	URL: Education Islamique 1	URL	Module de cours consulté	The user with id '174' viewed the 'url' activity with course	web	196.89.172.197
28 juin 20	LINA ELGRAOUI	-	Cours: EMA-1BAC	Système	Cours consulté	The user with id '174' viewed the course with id '59'.	web	196.89.172.197
28 juin 20	AYA BOUCHAMI	-	URL: Arabe	URL	Module de cours consulté	The user with id '231' viewed the 'url' activity with course	web	41.140.161.221
28 juin 20	IMANE ABAIK	-	Cours: EMA-1BAC	Système	Cours consulté	The user with id '220' viewed the course with id '59'.	web	196.217.60.205
28 juin 20	AYA BOUCHAMI	-	URL: Histoire	URL	Module de cours consulté	The user with id '231' viewed the 'url' activity with course	web	105.156.73.44
28 juin 20	AYA BOUCHAMI	-	Cours: EMA-1BAC	Système	Cours consulté	The user with id '231' viewed the course with id '59'.	web	105.156.73.44
28 juin 20	AYA BOUCHAMI	-	Cours: EMA-1BAC	Système	Cours consulté	The user with id '231' viewed the course with id '59'.	web	41.140.161.221
28 juin 20	AYA BOUCHAMI	-	URL: Histoire	URL	Module de cours consulté	The user with id '231' viewed the 'url' activity with course	web	41.140.161.221
28 juin 20	AYA BOUCHAMI	-	Cours: EMA-1BAC	Système	Cours consulté	The user with id '231' viewed the course with id '59'.	web	41.140.161.221
28 juin 20	AYA BOUCHAMI	-	URL: Géographie	URL	Module de cours consulté	The user with id '231' viewed the 'url' activity with course	web	41.140.161.221
28 juin 20	AYA BOUCHAMI	-	Cours: EMA-1BAC	Système	Cours consulté	The user with id '231' viewed the course with id '59'.	web	41.140.161.221
28 juin 20	AYA BOUCHAMI	-	URL: Education Islamique 1	URL	Module de cours consulté	The user with id '231' viewed the 'url' activity with course	web	41.140.161.221
27 juin 20	AYA YAAGOUBI	-	Cours: EMA-1BACSEF-1	Visites guidées	Visite guidée commencée	The user with id '182' has started the tour with id '1' on th	web	196.65.194.74
27 juin 20	AYA YAAGOUBI	-	Cours: EMA-1BACSEF-1	Système	Cours consulté	The user with id '182' viewed the course with id '22'.	web	196.65.194.74
26 juin 20	SOUHAIL BENHA	-	URL: Histoire	URL	Module de cours consulté	The user with id '229' viewed the 'url' activity with course	web	196.70.121.45

Fig. 1. Moodle Log File.



EMA-2BACSMF-A+B

Calculé à partir des journaux depuis mardi 19 mai 2020, 20:51.

Activité	Affichages	Articles de blogs en lien	Dernier accès
Annonces	-	-	
Salle virtuelle	3 consultations par 1 utilisateurs	-	lundi 15 juin 2020, 09:33 (13 jours 5 heures)
Physique			
La mécanique	27 consultations par 16 utilisateurs	-	dimanche 21 juin 2020, 06:34 (7 jours 8 heures)
La chute verticale d'un solide dans un fluide	13 consultations par 8 utilisateurs	-	dimanche 21 juin 2020, 06:34 (7 jours 8 heures)
Les ondes	8 consultations par 6 utilisateurs	-	jeudi 18 juin 2020, 18:33 (9 jours 20 heures)
Les transformations nucléaires	9 consultations par 5 utilisateurs	-	jeudi 18 juin 2020, 21:09 (9 jours 17 heures)
L'éléctricité	7 consultations par 7 utilisateurs	-	samedi 13 juin 2020, 13:31 (15 jours 1 heure)
Le dipôle RL	2 consultations par 2 utilisateurs	-	samedi 13 juin 2020, 13:44 (15 jours 1 heure)
La modulation et la démodulation d'amplitude	7 consultations par 5 utilisateurs	-	dimanche 21 juin 2020, 06:34 (7 jours 8 heures)

Fig. 2. Course Participation Report.

The LMS Moodle store a large number of information about teacher student interactions about access to resources and system, it produce many traces as low-level information. This information describes how and when students perform their missions, tasks and course engagement, but it is difficult for tutors to support learners in their educational course, because the information big volume it is difficult for tutors to deal with, which requires data collection and treatment in order to deliver meaningful information.

In this part, we place the traces cited in the literature and which appear among the factors influencing the learning process. We propose a model to analyze and predict students' learning performance based on three indicators; session indicator, productivity indicator and interactivity indicator. We collect data from Moodle and analyze it in order to be used to reveal knowledge and useful information from which decisions could be made by tutors in time.

These traces are: the degree of interest expressed by students to e-learning in general, the frequency of Internet connection per week, the frequency of consultation of the LMS per week, the average learning time, the average number of course units visited during a learning session, the degree of use of the chat and / or forum, the number of messages sent by chat and / or forum during a learning session, learners' productions and the score obtained in the course unit validation test.

We decide to predict the first activity indicator named session indicator based on the frequency of consultation of the courses on the platform per week, the average learning time, the average number of course units visited during a learning session. The second indicator is the productivity indicator is predicted from the number of learners' productions and the score obtained in the course unit validation test. The last indicator is the interactivity indicator is obtained based on the degree of use of the chat and / or forum, the number of messages sent by chat and / or forum during a learning session.

III. Establish a Fuzzy Logic System to Calculate Activity Indicators

Among the known techniques of artificial intelligence, we have chosen fuzzy logic because it is an approach based on human reasoning rather than on rigid calculations which is need in our study case. The fuzzy logic reasoning method is more intuitive and closer to human reason and thinking unlike classic logic. It allows a better understanding and interpretation of natural phenomena, thanks to the inference rules and the membership functions of fuzzy sets. These phenomena are often imprecise and difficult to model, which is applicable in our case where we measure the performance of learning activities in e-learning LMS.

A. *Introduction to Fuzzy Logic*

Nowadays, fuzzy logic is an important research area on which many scientists focus. Technological benefits are already available. The theoretical foundations of fuzzy logic were formulated in 1965 by Professor Lotfi A. Zadeh, of the University of Berkeley in California [13]. Fuzzy logic is based on fuzzy set theory, which is a generalization of classical set theory. By introducing the notion of degree in the verification of a condition, thus allowing a condition to be in a state other than true or false, fuzzy logic confers a very appreciable flexibility to the reasoning using it, making possible to take into account imprecisions and uncertainties [14].

1) Fuzzy sets and subsets: One of the interests of fuzzy logic to formalize human reasoning is that the rules are stated in natural language.

Definition: Let X be a set. A fuzzy subset A of X is characterized by a membership function fa: X → [0,1].

Note: this membership function is the equivalent of the characteristic function of a classic set [14].

According to the usual practices of the literature, the terms fuzzy subsets and fuzzy sets are used interchangeably. Classical sets are also called net sets, as opposed to fuzzy, and similarly classical logic is also called Boolean or Binary logic.

The Fig. 3 shows graphically the difference between a classic set and a fuzzy set.

2) Membership function: In our case, we take example of the first activity indicator, which is the session indicator; we will have to redefine membership functions for each fuzzy subset of each of our four variables:

Input 1: the frequency of consultation of the course. Subsets: low, medium and high.

Input 2: the duration of the learning session. Subsets: low, medium and high.

Input 3: the number of course units studied. Subsets: low, medium and high.

Output: the session indicator. Subsets: low, medium and high.

In a reference set X, a fuzzy subset A of this frame of reference is characterized by a membership function µ of A, which associates with each element x of X, the degree µA (x), between 0 and 1, for which x belongs to A. The value of the membership function at element x represents the "grade of membership" of x in A.

This function is the extension of the characteristic function of a classical subset. It can be represented as a Triangular, Gaussian, Trapezoidal or parabolic function. For the sake of clarity and to facilitate calculations, we will only use the first two forms.

We present a comparison between a characteristic function of a classical set and a membership function of a fuzzy set, in Fig. 4.

Fuzzy logic is applied in many fields (industrial processes, commerce, chemical industry, etc.) where it gives very satisfactory results. In our case, it is applied to the prediction of indicators of learner activity.

The fuzzy logic system is built by choosing independent variables that describe well the dependent variable. Fuzzy sets in human language are used to describe a variable instead of using the numeric value. The membership function determines the degree of certainty that each variable belongs to a fuzzy set. The inference engine is responsible for applying each of the inference rules. They represent the knowledge that we have of the system due to human expertise. Each rule will generate an exit command. Then comes the step allowing merging different commands generated by the inference engine to give only one output command and to transform this linguistic output variable into numerical data.

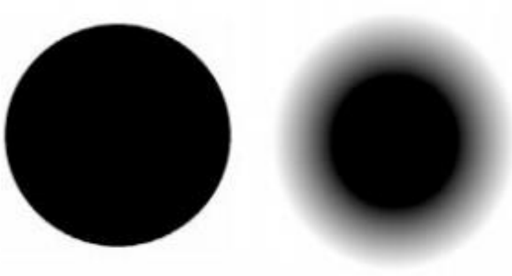


Fig. 3. Comparing Graphical Representations of Classical and Fuzzy Sets.

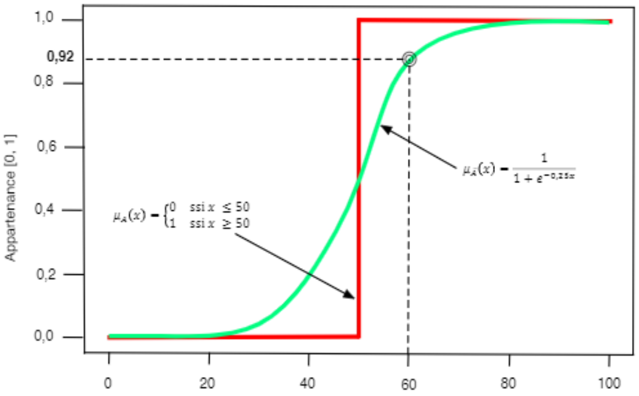


Fig. 4. Comparing Membership Functions of Fuzzy and Classic Sets.

B. Fuzzy Logic applied to Predict Activity Indicators

1) Data sources and simulation software: For the realization of this model we used data collected from the Learning Management System Moodle, this information on the secondary students in their last year as well as on their history of login files and course participation reports. We used Matlab for modeling the fuzzy system. We capture all input membership functions and the output function. The inference rules are also captured.

2) Session indicator fuzzy inference system: The first activity indicator named Session indicator is based on the frequency of consultation of the courses on the platform per week “freqCons”, the average learning time ”dureeSess”, the average number of course units visited during a learning session “nbrUnit”; it is represented in Fig. 5.

The inputs variables of the Session indicator fuzzy system are the freqCons, dureeSess and the nbrUnit. After the process of Fuzzification and Defuzzification we obtain the output variable: “IndicSess”.

3) Productivity indicator fuzzy inference system: The second indicator is the productivity indicator, it is the result of the number of learners' productions and the score obtained in the course unit validation test; it is represented in the Fig. 6.

The inputs variables of the productivity indicator fuzzy system are the production and the score. After the process of Fuzzification and Defuzzification we obtain the output variable: “IndicProd”.

4) Interactivity indicator fuzzy inference system: The last indicator is the interactivity indicator, it is the result of the degree of use of the chat and / or forum, the number of messages sent by chat and / or forum during a learning session; it is represented in the Fig. 7.

The inputs variables of the productivity indicator fuzzy system are the nbrMsg and the ChatForum. After the process of Fuzzification and Defuzzification we obtain the output variable: “IndicInterac”.

5) Fuzzification and defuzzification: The purpose of the fuzzification step is to transform a digital data into a linguistic variable; the inference mechanism used is based on the Mamdani model [13]. In the second step, we move on to define the linguistic variables and their membership functions. We used two types of membership functions for inputs and outputs because each variable has its own characteristics and can be modeled differently from another. The last step is the defuzzification. In the second step, we generated a set of commands in the form of linguistic variables (one command per rule). The purpose of defuzzification is to merge these controls and transform the resulting parameters into numeric data.

The aim of this paper is to calculate learners’ activity indicators in e-learning using a fuzzy logic system. By modifying the entries for this system the indicators varies accordingly.

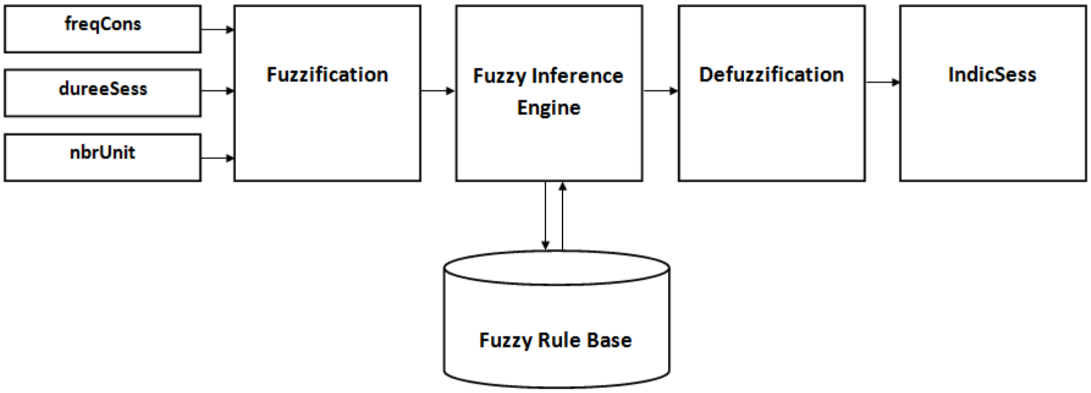


Fig. 5. The Architecture of the Session Indicator using Fuzzy Logic System Architecture.

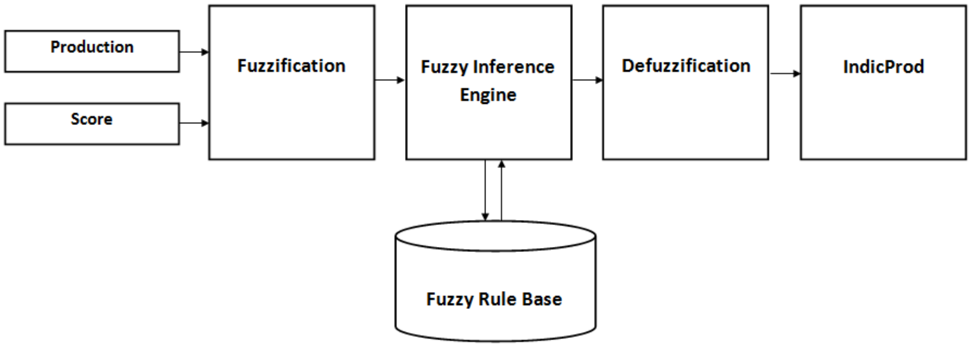


Fig. 6. The Architecture of the Productivity Indicator using Fuzzy Logic.

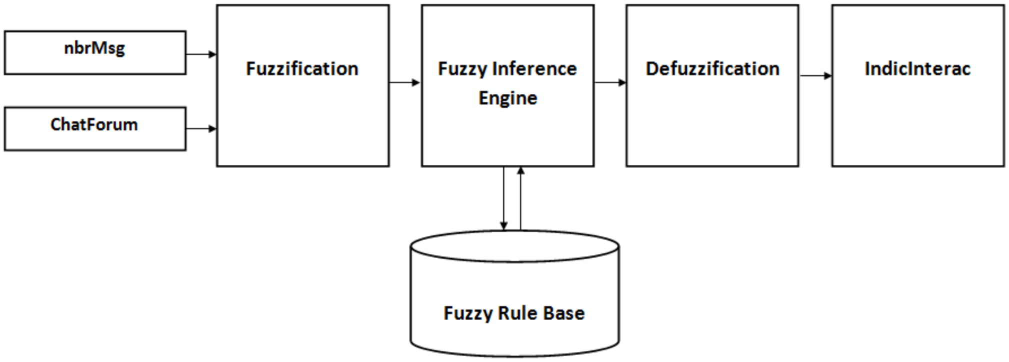


Fig. 7. The Architecture of the Interactivity Indicator using Fuzzy Logic.

IV. Implementation of Activity Indicators Fuzzy Logic Motivation System

The use of linguistic variables makes it possible to introduce great flexibility into the characterization of fuzzy, vague or imprecise descriptions, and to avoid the artificially rigid boundaries of standard descriptions (statistical or probabilistic descriptions). The fuzzy model therefore makes it possible to be closer to reality.

When two descriptions are close, or when they can be deduced from each other, this leads to modifiers such as "very", "more or less", which constitute a kind of standardization of the linguistic expression used to modulate a description.

A. Simulation Environment

MATLAB is an easy and efficient programming environment; it constitutes an interactive and user-friendly system of numerical calculation and graphic visualization. Aimed at engineers and scientists, it is a widely used tool, in universities and in the industrial world. Matlab integrates hundreds of mathematical and numerical analysis functions (matrix calculation, signal processing, image processing, graphic visualizations, etc.).

B. Fuzzy Logic Indicators Simulation

The purpose of this research is to predict the learning activity indicators of learners in e-learning. We got the idea to use a fuzzy expert system for the first time in learning activity indicators, considering that these indicators can be expressed using human reasoning language. Suppose that it can be evaluated according to a scale of appreciation, which to be easily usable by the person who evaluates the item, must be linguistic. Indeed, it is more natural to express an assessment in the form "the learner is rather producer" than to give a grade, whatever the scale used.

The fuzzy expert systems of the three activity indicators calculated based on traces collected from the LMS Moodle. By modifying the entries for these systems the indicators varies accordingly.

The membership functions of each evaluation are represented by a triangle, a Gaussian curve or a trapezoid, whose base or large base, respectively, covers part of the symbolic scale (in our case, [0.4]). The equation (1) of the Triangular membership functions of the freqCons and the nbrUnit variables:

$$\mu A(x) = \begin{Bmatrix} \frac{x - x1}{x2 - x1} \text{ si } x \in [x1\,, x2] \\ \frac{x3\ - x}{x3 - x2} \text{ si } x \in [x2\,, x3] \end{Bmatrix} \tag{1}$$

The Triangular membership functions of the variables of the fuzzy activity indicators are represented in Fig. 8 and 9.

The equation (2) of the Gaussian membership functions of the dureSess and the IndicSess variables:

$$\mu A(x) = \begin{Bmatrix} e^{-\frac{(x-c)2}{2\sigma g2}}\ si\ x < c \\ e^{-\frac{(x-c)2}{2\sigma d2}}\ si\ x\ \geq c \end{Bmatrix} \tag{2}$$

The Gaussian membership function of the variable dureSess and the output variable of the first activity indicator are represented in Fig. 10 and 11.

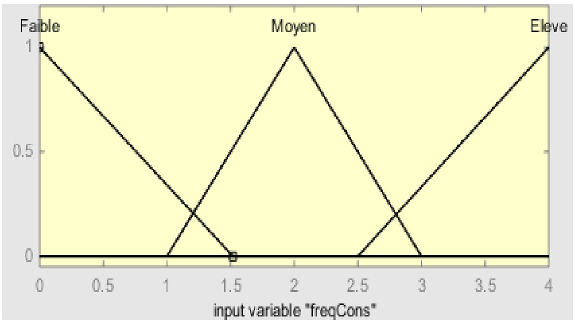


Fig. 8. The Membership Function of the First Input Variable FreqCons.

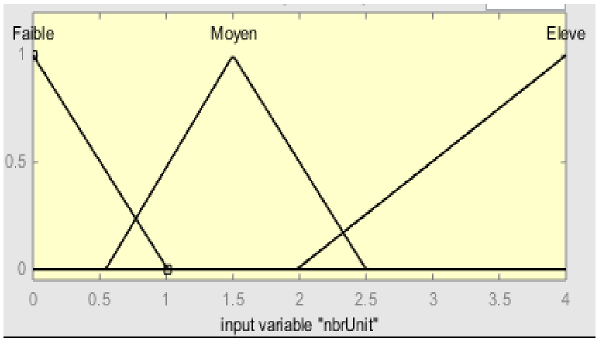


Fig. 9. The Membership Function of the Second Input Variable nbrUnit.

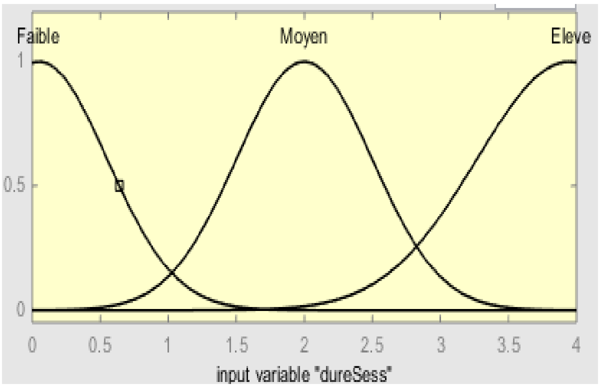


Fig. 10. The Membership Function of the Third Variable DureSess.

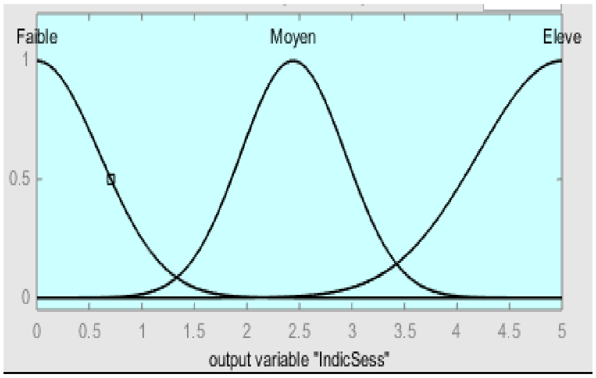


Fig. 11. The Membership Function of the Output Variable IndicSess.

The inference rules of the Session Indicator fuzzy model Fig. 12.

The 3D surface simulations of the Session indicator according to the input variables are presented in the Fig. 13, 14 and 15.

Concerning the membership functions of the inputs variable of the productivity indicator, we used the Triangular membership function for the variable score and the Gaussian one for the variable production and the output variable IndicProd.

The 3D surface simulations of the productivity indicator in function of the inputs variables are presented in Fig. 16.

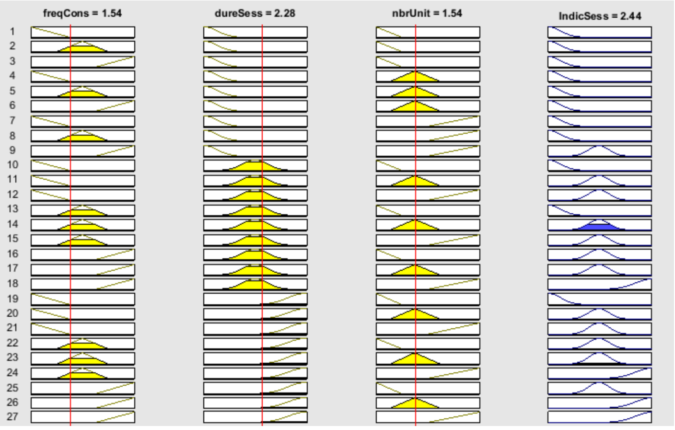


Fig. 12. The Inference Rules of the Session Indicator.

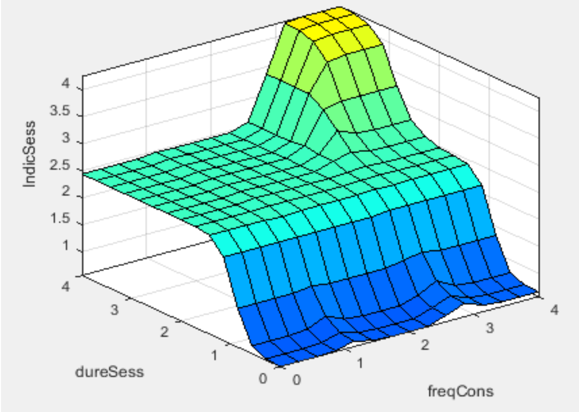


Fig. 13. 3D Surface Simulation of Session Indicator in Function of the Variables DureeSess and FreqCons.

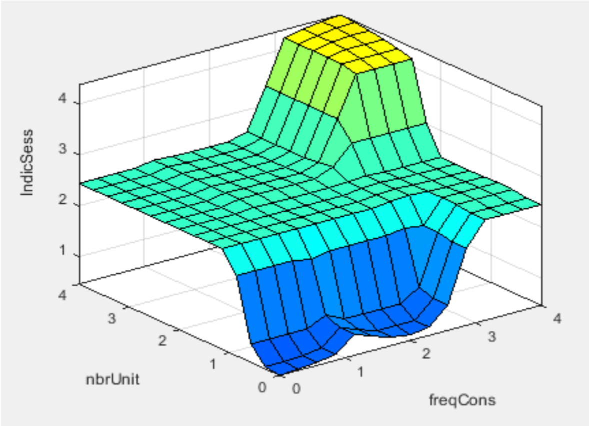


Fig. 14. 3D Surface Simulation of Session Indicator in Function of Variables NbrUnit and FreqCons.

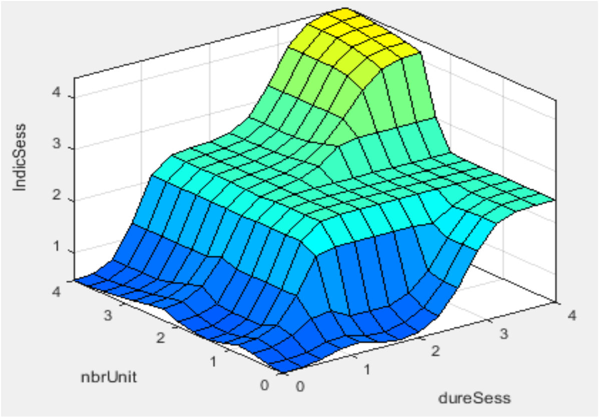


Fig. 15. 3D Surface Simulation of Session Indicator in Function of NbrUnit and DureSess Variables.

The inputs and outputs variables for the Interactivity indicator are presented with a Triangular and Gaussian membership function. The Fig. 17 presents the 3D simulation of the Interactivity indicator.

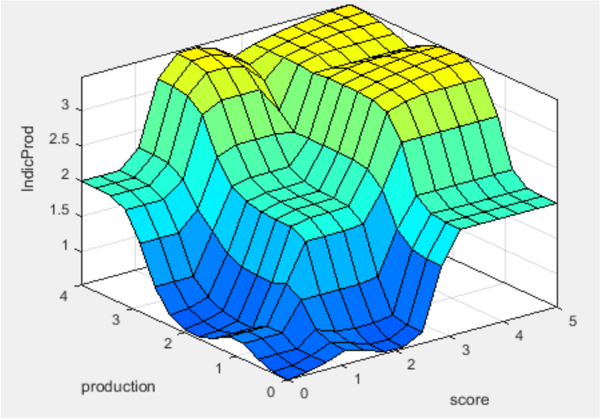


Fig. 16. 3D Surface Simulation of Productivity Indicator in Function of Production and Score Variables.

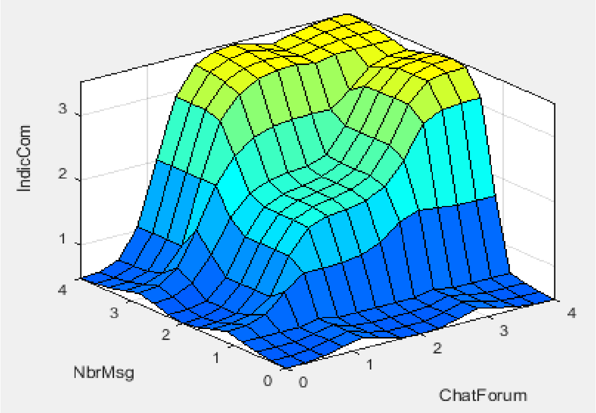


Fig. 17. 3D Surface Simulation of Interactivity Indicator in Function of Variables ChatForum and NbrMsg.

C. Discussion

In this contribution, we presented our innovative idea to elaborate a system for learners' activity indicators prediction based on fuzzy logic; it's the first time that fuzzy logic is used in learning activity prediction in order to help the tutor in decision-making when tracking learners in e-learning. Thus, fuzzy logic makes it possible to set up inference systems whose decisions are non-discontinuous, flexible and non-linear. It is closer to human reasoning unlike classical logic, because the rules are written in natural language.

Through this research paper, we have shown the activity indicators calculated from the traces of learners collected from the LMS Moodle in the learning performance. In addition, this clearly showed that the integration of fuzzy logic significantly improves tutor's work and gives him a clear vision of the learning process.

We could see that the value of the three indicators has a strong relationship with the values of the traces collected from Moodle, because it is clear that the value of each indicator is high when the values of the variables are high and it is low otherwise.

V. Conclusion

We started the research work with an inventory of the problems encountered when using e-learning. During a learning activity, learners leave traces stored and recorded in dedicated files of the LMS. Until now, this data is stored as low level information with a large volume requiring processing. The task of collection, treatment and analyzing this data to extract meaningful information is very heavy. In addition, virtual mode does not offer the same working conditions as in face-to-face mode.

To remedy this big problem, we decided to set up a tracking system for learning activity, by collecting information from the LMS Moodle which are called traces, after we passed to the filtering of this information to keep only the most relevant which better describe the course of the learner's activity. In the next step we chose to elaborate the activity indicators prediction system using fuzzy logic. These activity indicators give a clear idea to tutors of how the course is learned based on meaningful information about learners, it could give tutors a high level processing of the learning activity.

After having almost wiped out all the studies done on the level of learner motivation, we have found that it plays a key role in preserving learners' persistence, which also guarantees the continuity of training.

Once we got to this stage, we created a motivation prediction system using fuzzy logic, since motivation is a rough concept that cannot be measured. This decision-making system will help the tutor to have a clear view of the e-learning process so that they can act accordingly.

Acknowledgment

We would like to warmly thank the members of our research team and the journal scientific committee that will evaluate our paper.

References

[1] A. F. Agudo-Peregrina, S. Iglesias-Pradas, M. Á. Conde-González, and Á. Hernández-García, "Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning". Computers in human behavior, 2014, vol. 31, pp. 542–550.

[2] M. Romero, "Analyser les apprentissages à partir des traces : Des opportunités aux enjeux éthiques". Distances et médiations des savoirs, 2019, vol. 26.

[3] M. Á. Conde, Á. Hérnandez-García, F. J. García-Peñalvo, and M. L. Séin-Echaluce, "Exploring Student Interactions: Learning Analytics Tools for Student Tracking". In: Zaphiris P., Ioannou A. (eds) Learning and Collaboration Technologies. LCT 2015. Lecture Notes in Computer Science, 2015, 9192, 50-61, Springer, Cham.

[4] D. Peraya, "Les Learning Analytics en question". Distances et médiations des savoirs, 2019, vol. 25.

[5] S. Chehbi, R. Elouahbi, and F. El Khoukhi, "Collection and transformation of Moodle traces". In 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), pp. 570–574.

[6] J. Laflaquière, J., L. S. Settouti, Y. Prié, and A. Mille, "Traces et inscriptions de connaissance", 18es Journées Francophones d'Ingénierie des Connaissances, 2007, Grenoble, France.

[7] B. Florian, C. Glahn, H. Drachsler, M. Specht, and R. Fabregat Gesa, "Activity-Based Learner-Models for Learner Monitoring and Recommendations in Moodle". In: Kloos C.D., Gillet D., Crespo García R.M., Wild F., Wolpers M. (eds) Towards Ubiquitous Learning. EC-TEL 2011. Lecture Notes in Computer Science, 6964. Springer, Berlin, Heidelberg.

[8] S. Iglesias-Pradas, C. Ruiz-de-Azcárate, Á. F. Agudo-Peregrina, “Assessing the suitability of student interactions from Moodle data logs as predictors of cross-curricular competencies”. Computers in Human Behavior, 2015, vol. 47, pp. 81–89.

[9] T. Djouad, and A. Mille, “Observing and understanding an on-line learning activity: A model-based approach for activity indicator engineering”. Technology, Knowledge and Learning, 2018, vol. 23(1), pp. 41–64.

[10] E. A. Gamie, M. S. A. El-Seoud, M. A. Salama, and W. Hussein, “Pedagogical and Elearning Logs Analyses to Enhance Students' Performance”. In Proceedings of the 7th International Conference on Software and Information Engineering, 2018, pp. 116–120.

[11] Á. Hernández-García, E. Acquila-Natale, S. Iglesías-Pradas, and J. Chaparro-Peláez, “Design of an Extraction, Transform and Load Process for Calculation of Teamwork Indicators in Moodle”. In LASI-SPAIN, 2018, pp. 62–73.

[12] R. Mazza, M. Bettoni, M. Faré, and L. Mazzola, “Moclog–monitoring online courses with log data”. 2012.

[13] L. A. Zadeh, “Soft computing and fuzzy logic”. In Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems. Selected Papers by Lotfi a Zadeh, 1996, pp. 796–804.

[14] F. Dernoncourt, and E. Métais, “La Logique Floue: le raisonnement humain au cœur du systeme décisionnel? ”. Memory NFE211 engineering decision systems Paris, February 2011.

An Evolutionary Algorithm for Short Addition Chains

Hazem M. Bahig[1]*, Khaled A. Alutaibi[2], Mohammed A. Mahdi[3], Amer AlGhadhban[4], Hatem M. Bahig[5]
Computer Science and Information Department, College of Computer Science and Engineering, University of Ha'il, Ha'il, KSA[1,3]
Computer Engineering Department, College of Computer Science and Engineering, University of Ha'il, Ha'il, KSA[2]
Electrical Engineering Department, College of Engineering, University of Ha'il, Ha'il, KSA[4]
Computer Science Division, Mathematics Department, Faculty of Science, Ain Shams University, Cairo, Egypt[1, 5]

***Abstract*—The encryption efficiency of the Rivest-Shamir-Adleman cryptosystem is based on decreasing the number of multiplications in the modular exponentiation (ME) operation. An addition chain (AC) is one of the strategies used to reduce the time consumed by ME through generating a shortest/short chain. Due to the non-polynomial time required for generating a shortest AC, several algorithms have been proposed to find a short AC in a faster time. In this paper, we use the evolutionary algorithm (EA) to find a short AC for a natural number. We discuss and present the role of every component of the EA, including the population, mutation operator, and survivor selection. Then we study, practically, the effectiveness of the proposed method in terms of the length of chain it generates by comparing it with three kinds of algorithms: (1) exact, (2) non-exact deterministic, and (3) non-exact non-deterministic. The experiment is conducted on all natural numbers that have 10, 11, 12, 13, and 14 bits. The results demonstrate that the proposed algorithm has good performance compared to the other three types of algorithms.**

Keywords—Addition chain; short chain; evolutionary algorithm; modular exponentiation; RSA

I. Introduction

The purpose of cryptography is to secure communications over an open channel. To achieve this, two procedures are used, encryption and decryption. In simple terms, encryption is a process of transforming a given data, called plaintext, to unmeaningful data, called ciphertext, which can be reinstated to original form through the process of decryption.

One of the modern and strong encryption techniques used to encrypt data is the Rivest-Shamir-Adleman (RSA) cryptosystem. Encryption in RSA is based on generating a secret message c from the original message m using the modular exponentiation (ME) formula.

$c \equiv m^e \bmod N$

where the pair (e,N) is the public key and N is a composite odd number and equal to the product of two prime numbers. Similarly, for decryption, one needs to compute.

$m \equiv c^d \bmod N$

where d is the private key.

The main challenge associated with the encryption and decryption techniques in RSA is that the time consumed, in particular for decryption, is quite large because its performance is based on the ME mathematical operation that requires high computation resource. The time consumed in ME is due to the computation of a sequence (thousands, 1024-4096) of multiplications and squares for large-size numbers. Therefore, several techniques have been proposed to speed up the computation of multiplication and ME using sequential and parallel computation, for examples [6, 8, 19, 20, 35].

One of the strategies that can be used to decrease the number of multiplications in ME involves the use of an addition chain (AC) [27, 30]. An AC [22] for a number e is a finite increasing sequence of non-repeated natural numbers such that the start element of the sequence is 1 and any next element in the sequence is the sum of any two previous elements that are not necessarily distinct. The final element in the sequence is e. There are different types of chain besides the ACs, such as addition-subtraction chains [21,28], addition-multiplication chains [2, 9], Lucas chains and q-chains [22, 25] and addition sequence [7,17].

For a natural number e, there are many ACs of different lengths. A chain with minimal length is called a shortest AC, otherwise it is called a short AC. Therefore, if we can construct a shortest AC for e as $e_0 = 1, e_1 = 2, e_2, \dots, e_\ell = e$, then m^e can be computed in a minimal number of multiplications as m, m^2, m^{e_2},…, , m^e.

For example, for the natural number $e = 37$, we can construct different chains for e with different lengths as (1) $C_1 = (1,2,3,6,9,15,17,20,29,35,37)$ of length 10, (2) $C_2 = (1,2,3,5,10,20,30,35,37)$ of length 8, and (3) $C_3 = (1,2,4,8,16,32,36,37)$ of length 7. Therefore, the minimum number of multiplications to compute m^{37} is 7 and can be computed as m, $m^2 = m \times m$, $m^4 = m^2 \times m^2$, $m^8 = m^4 \times m^4$, $m^{16} = m^8 \times m^8$, $m^{32} = m^{16} \times m^{16}$, $m^{36} = m^{32} \times m^4$, $m^{36} = m^{32} \times m^4$, $m^{37} = m^{36} \times m^1$.

A large number of algorithms have been designed to find a solution to the AC problem. The goal of these algorithms is either to (1) find a shortest AC or (2) find a short AC. In respect of the first goal, the branch and bound method is one of the efficient strategies that has been used to find a shortest AC (see for examples, [1,3,12,34]). Another improvement strategy that has been suggested to find a shortest AC is to use high-performance computing to speed up the computation required to find the exact solution. The authors in [5] and [10] used a graphics processing unit and multicore system, respectively, to find a solution for the AC more quickly as compared to using sequential algorithms. However, while these solutions produce a chain that is shortest and therefore the solution is optimal, the running time of these algorithms is in non-polynomial time.

*Corresponding Author

On the other hand, many non-exact algorithms [4,13-16,18,23,24,26-33,36] have been proposed to find a short AC. In these algorithms, the length of the generated AC is not necessarily minimal and the algorithms run in polynomial time. Many different strategies have been proposed to find a solution to produce a short AC. We can classify these algorithms into two classes based on their behavior into deterministic algorithms and non-deterministic algorithms. The non-deterministic algorithm may generate different lengths in different runs for the same input e, while the deterministic algorithm produces the same length of AC even with different runs.

Examples of deterministic algorithms [4,11,18,20,22,30] for a short AC include the binary method, window method, factor method, and continued fraction method. The difference between these methods lies in the strategy used to find the solution, the length of the generated short AC, and the running time. In general, the running time of these algorithms is very fast when compared to that of exact algorithms. Also, the window and continued fraction methods give a better output than other methods.

Examples of non-deterministic algorithms [13-16,23,26-29,31-33] for a short AC include the genetic algorithm (GA), evolutionary algorithm (EA), ant colony algorithm, swarm intelligence algorithms, and artificial immune algorithm. All these algorithms are based on many factors such as the size of the population, the maximum number of iterations, and the strategies of different operators such as crossover and/or mutation. In general, non-deterministic algorithms can produce a short AC with a length that is better than that generated by the deterministic algorithms, see [26, 32].

From a review of the previous works related to the generation of a short AC, we can make several of observations. First, experimental studies have been conducted on (1) certain numbers only or (2) a small set of numbers within a data range such as 30 random numbers in the range [1,2048]. Second, in some research works, the value of some parameters, such as the total number of generations, is large. This leads to an increase in the running time of artificial intelligence (AI) algorithms. Third, in some previous algorithms, the number of independent runs for each natural number is 30 or 50, which is a large number and leads to increase running time.

In this paper, we are interested in using the EA to find an approximate solution for the AC problem. The proposed algorithm is based on modifying one of the EAs [16] for a short AC in relation to three aspects (1) the process of generating the elements of chain, (2) mutation operator, and (3) survivor selection. The developed algorithm exhibits good performance in terms of the generated length of the AC as compared to previous methods. To prove the effectiveness of the developed algorithm, we conducted an experimental study on all integers with 10, 11, 12, 13, and 14 bits. We used two types of algorithms for the comparison: (1) an exact algorithm that gives a minimal length AC; and (2) non-exact algorithms (deterministic and non-deterministic) that give a short chain.

The rest of the paper consists of five sections. In Section II, we describe the background to the AC problem and the EA. Next, in Section III, we briefly review the related works on the AC problem that use AI strategies to find a short chain. Then, in Section IV, we present the proposed algorithm and the details of the EA for a short AC. This is followed by Section V, in which we provide the results of our measurements to determine the effectiveness of the proposed algorithm on different ranges of bits for the exponent e, as well as the results of a comparison with the outputs of exact and approximate solutions in the literature. Finally, in Section VI, we conclude the work and highlight our future works.

II. Background

In this section, first, we provide a brief background on the definition of the AC problem, the types of chain and their elements [22]. Then, we give an overview of the EA and its components.

A. The Addition Chain Concept

An AC for a natural number e is a sequence of natural numbers e_0,e_1,e_2,…,e_m such that.

- $e_0 = 1$,
- $e_i = e_j+e_k$, $0\leq k \leq j < i$, and
- $e_m = e$.

The length of the AC for e is the number of steps needed to compute e, which is equal to the number of elements in the chain minus one, m. The second rule for generating a chain for e leads to the possibility of constructing a large number of different chains for e. This means that the lengths of the different chains for the same e may be different in general. Therefore, there are two types of chain based on the length of the chain as follows:

- If the length, m, of an AC for e is minimal, the length of the chain is denoted as $\ell(e)$ and the chain is termed a shortest AC.
- If the length, m, of the AC for e is greater than $\ell(e)$, the chain is termed a short AC.

For the natural number e, we define two functions as follows:

- $\lambda(e)$ is the length of the binary representation for e minus one and equal to $\lfloor log_2 e \rfloor$.
- $v(e)$ is the number of 1's in the binary representation of e.

There are different types of steps involved in the generation of an AC. The important steps are defined as follows:

- The ith step in an AC is termed star, if $e_i = e_{i-1}+e_k$, $0\leq k < i$.
- The ith step in an AC is termed doubling, if $e_i = e_{i-1}+e_{i-1}$.
- The ith step in an AC is termed plus-one, if $e_i = e_{i-1}+e_0$.

B. The Evolutionary Algorithm

An EA is a generic population-based metaheuristic optimization algorithm. An EA uses mechanisms inspired by biological evolution, such as reproduction, mutation, recombination, and selection.

Given a quality function to be maximized/minimized, a set of candidate solutions or population, called parents (i.e., elements of the function's domain), can be randomly generated. Then, a quality function can be applied to these candidates to evaluate their fitness values, where the higher the fitness the better. Based on these fitness values, the best candidates are chosen to seed the next generation. This is done using two main strategies:

- A variation operator that generates the necessary variety within the population to be used in the next generation. Examples of variation operators are recombination and mutation.
- Selection that acts as a force that increases the mean quality of the solutions in the population.

Recombination is an operator that is applied to two or more selected candidates to produce one or more new candidates (children). One the other hand, the mutation operator uses one candidate, which results in one new candidate. Hence, these operations on candidates to create a set of new candidates (offspring). The candidates' fitness levels are evaluated before they compete – based on their fitness (and sometimes age) – with the old leads for a place in the next generation. This process is repeated until either a candidate with sufficient quality is found or a previously set computational limit is reached. Note that a combination of variation and selection leads to improved fitness values of consecutive populations.

III. Related Works

In this section, we briefly review different AI techniques that have been proposed to find a short AC.

A number of researchers have presented solutions to the AC problem based on the GA. Nedjah and Mourelle [27,28] used binary encoding to represent the solution, where 1 indicates that the number is present in the AC and 0 otherwise. They used four standard crossover operations: single-point, double-point, uniform, and arithmetic. Mutation is done by randomly changing some genes from 0 to 1 and vice-versa. The fitness function is based on the validity of the AC and its length. Cruz-Cortés et al. [13] adopted an integer encoding approach using variable-length chromosomes. They use a one-point crossover operator and the fitness function is based on the length of the AC.

The GA proposed by Osorio-Hernández et al. [31] works only on valid ACs (invalid chains are discarded) and represents each number from the AC corresponding to a gene in the chromosome. They use a repair process to generate valid solutions in the initial population and mutation operator. Also, they use a two-point crossover operator which applies value and rule copying operations. The fitness function is again based on the length of the AC.

Domínguez-Isidro et al. [16] proposed using an EA, which is based on a mutation operator that is able to produce as set of valid solutions to the AC problem from a single solution. In addition, the proposed algorithm includes a replacement technique based on stochastic elements to introduce diversity into the population. The numbers in the AC are represented directly in the solution and their fitness is the length of the AC.

Picek et al. [32] presented a GA with a repair strategy to enhance the performance of AC generation. The solution is encoded as a set of tuples of the form (v_k,i,j), where v_k is the value of the kth number in the AC while i and j are the positions of the previous numbers v_i and v_j, respectively, in the AC forming v_k. The algorithm implements crossover and mutation operators similar to the ones used in the previous literature [27,28]. However, these operators are followed by a repair operation to guarantee the validity of the solution.

Other researchers have used optimization techniques to find short ACs. Nedjah and Mourelle [29] proposed an ant colony optimization (ACO) approach based on a multi-agent schema with two types of memory: shared and local. The same authors implemented the ACO algorithm on a system-on-chip (SoC) to improve the computations [30].

On the other hand, Léon-Javier et al. [23] proposed algorithm based on particle swarm optimization (PSO). Mullai and Mani [26] used PSO and simplified swarm optimization to generate ACs for RSA for the purpose of using ACs to optimize computations in encryption/decryption processes to reduce processing time and power consumption in mobile devices. Cruz-Cortes et al. [14] introduced an artificial immune system for finding short ACs for moderate-sized exponents (i.e., less than 20 bits) and large exponents (i.e., up to 2048 bits).

IV. Proposed Algorithm

In this section, we describe the main steps of the proposed algorithm that is aimed at solving the AC problem by using an EA. The input of the algorithm is a natural number $e \geq 2$ and the output is a short AC $e_0, e_1, e_2, \ldots, e_m$.

In order to describe the proposed algorithm using EA, first, we describe the main components of EA involved in solving the AC problem, namely, representation, initial population, fitness function, variant operators, and survivor selection. Then, we present the steps of the proposed algorithm.

A. Representation

Since an AC for a natural number e is a sequence of natural numbers, $e_0, e_1, e_2, \ldots, e_m$, the individual AC in the proposed EA is an array of dynamic length. The length of each individual is not fixed during the computation for two reasons. The first reason is the length of each individual in the search space may be different from the others. The second reason is the length of an individual AC may change during the different mutation operations.

In order to represent the population of the problem, we assume that the number of elements in the population for the AC problem is n. The population represents as a 2-dimensional array, E. The first dimension represents the number of individuals in the search space, n, while the second dimension is the length of each individual which is variable.

For example, let us assume that e = 37 and that we have four ACs for e, (1,2,4,8,16,32,36,37), (1,2,3,6,9,15,17,20,29,35,37), (1,2,4,8,16,18, 36,37), and (1,2,3,5,10,20,30,35,37). Then the four chains can be represented as four arrays of different lengths, i.e., 8, 11, 7, and 8, respectively, as in Fig. 1.

1	1	2	4	8	16	32	36	37			
2	1	2	3	6	9	15	17	20	29	35	37
3	1	2	4	8	16	18	36	37			
4	1	2	3	5	10	20	30	35	37		

Fig. 1. Four Chains for *e*=37.

B. Evaluation Function

Each individual AC needs to be evaluated using a fitness function to decide the best set of solutions. Since the goal of solving the AC problem is to find a sequence of natural numbers with minimal length, the value of the fitness function for each individual is the length of chain. Formally, if we have a chain $E_k = (e_{k,0}, e_{k,1}, \ldots, e_{k,m})$, then $f(E_k) = m$, because the first element, e_0, in the chain is not counted. Note that, in some times, we write $E_k = (e_0, e_1, \ldots, e_m)$ for simplicity.

C. Initial Population

The first step in any EA is to generate a random initial population that consists of n individuals, i.e., ACs. Each individual should satisfy the following general conditions to be a valid AC:

- The first two elements in any chain are 1 and 2.
- Any element, e_i, in the chain can be constructed from the addition of any two previous elements, e_j and e_k, where $j, k < i$.
- No elements in the chain are repeated.
- The last element in the chain is *e*.
- The elements of the chain are in increasing order of size.

All elements in the AC, except the first two elements, are generated randomly. In order to generate the elements, we use the following variables:

- r_1, which is a real random number between 0 and 1. The variable is generated using a user-defined function *RandomReal*(r_1).
- r_2, which is an integer random number between 0 and an integer α-1. The variable is generated using a user-defined function *RandomInt*(r_2,α).
- *diff*, which is the difference between *e* and the current element, e_i, in the chain.

The previous proposed strategy [16] used to generate an element in a chain is based on applying one of the following rules based on a random number:

R1. Double the last element: $e_{i+1} = e_i + e_i$.

R2. Add the last two elements: $e_{i+1} = e_i + e_{i-1}$

R3. Add the last element with a random element from 0 to *i*-1: $e_{i+1} = e_i + e_j$, $0 \le j < i$.

In our proposed algorithm, we modify this strategy by measuring, first, the difference, *diff*, between the target number, *e*, and the last element generated in the chain, e_i. Therefore, we have three cases for the value of *diff* as follows:

Case 1: When *diff* is greater than or equal to e_i, the algorithm executes one of the rules, R1, R2 or R3, based on the value of a random number.

Case 2: When *diff* is greater than or equal to e_{i-1}, the algorithm executes one of the rules, R2 or R3, based on the value of a random number.

Case 3: When *diff* is less than e_{i-1}, the algorithm executes R3 based on the value of a random number.

One of the advantages of using *diff*, in the first two cases, compared to other previous strategies is that it prevents the occurrence of the following cases: (1) generation of repeated elements, (2) generation of an element greater than the target element, and (3) generation of a non-increasing sequence without making a repair to the chain. For the third case, we need to repair the generated elements such that the new element is less than or equal to *e*. Note that we can add more cases for the variable *diff*, but this leads to an increase in the running time in general.

The two subroutines InitialPopulation and GenerateSubChain are used to create the initial population.

```
Subroutine InitialPopulation($e, n$)
Begin
  For $k = 1$ to $n$ do
      $e_{k,0} = 1, e_{k,1} = 2$
      RandomReal($r_1$)
      If ($r_1 \ge 0.5$)
          $e_{k,2} = 4$
      Else
          $e_{k,2} = 3$
      $i = 2$
      GenerateSubChain($e$,$E_k$, $i$)
  Return $E$
End
```

```
Subroutine GenerateSubChain($e, E_k, i$)
Begin
    Repeat
      $diff = e - e_{k,i}$
      If ($diff \geq e_{k,i}$ ) then
        $RandomReal(r_1)$
        If ($r_1 \geq 0.5$)
          $e_{k,i+1} = e_{k,i} + e_{k,i}$
        Else
          $RandomReal(r_1)$
          If ($r_1 \geq 0.5$)
            $e_{k,i+1} = e_{k,i} + e_{k,i-1}$
          Else
            $RandomInt(r_2, i-1)$
            $e_{k,i+1} = e_{k,i} + e_{k,r_2}$
      Else
        If ($diff \geq e_{i-1}$ ) then
          $RandomReal(r_1)$
          If ($r_1 \geq 0.5$)
            $e_{k,i+1} = e_{k,i} + e_{k,i-1}$
          Else
            $RandomInt(r_2, i-1)$
            $e_{k,i+1} = e_{k,i} + e_{k,r_2}$
        Else
          $RandomInt(r_2, i-1)$
          $e_{k,i+1} = e_{k,i} + e_{k,r_2}$
          $j$=$i$-2
      While ($e_{k,i+1} > e$ do
        $e_{k,i+1} = e_{k,i} + e_{k,j}$
        $j$=$j$-1
      $i = i + 1$
    Until $e_{k,i} = e$
End
```

D. Mutation

In the EA, random strategy is used to mutate an individual element, AC, which means that for a given chain, E_k, we mutate E_k from a random position in the chain. This strategy to mutate an AC of length m is based on the following idea [16]: First, the algorithm picks, randomly, a position, j, in the AC between 3 and m. Second, the algorithm eliminates the elements of the AC from j to m and generates new random elements in the AC using the same strategy, i.e., the *GenerateSubChain* subroutine. Third, the algorithm repeats the second step t times and then the algorithm selects the best AC that has the smallest length.

In our algorithm, we modify the above method to mutate the AC so that the selected random position that is used to mutate the chain is different, if possible, in each iteration, where we have t iterations. In the above-described mutated strategy, the position of mutation is fixed during all t iterations. Also, when we generate a random position in the chain E_k, in our method, we select it from a range of 3 to the number of bits in e, say $\lambda(e)$.

In order to generate t mutated chains from the chain, E_k, we use two auxiliary arrays. The first auxiliary array, Aux_1, is used to save the best mutated chain from e, while the second auxiliary array, Aux_2, is used to generate the mutated chain from the chain, E_k. At the end of t iterations, the best mutated AC for e is selected as offspring.

```
Subroutine Mutation($e, E, n, t$)
Begin
    For $k = 1$ to $n$ do
        $RandomInt(r_2, \lambda(e) - 2)$
        $r_2 = r_2 + 2$
        $Aux_1(0, \dots, r_2) = E_k(e_0, \dots, e_{r_2})$
        $GenerateSubChain(e, Aux_1, r_2)$
        For $j = 2$ to $t$ do
            $RandomInt(r_2, \lambda(e) - 2)$
            $r_2 = r_2 + 2$
            $Aux_2(0, \dots, r_2) = E_k(e_0, \dots, e_{r_2})$
            $GenerateSubChain(e, Aux_2, r_2)$
            If ($|Aux_2| < |Aux_1|$) then
                $Aux_1 = Aux_2$
        $M_k = Aux_1$
    Return $M$
End
```

Fig. 2 shows an example of applying the mutation to the chain E_k = (1,2,3,6,9,15,17,20,29,35,37), where the four random numbers used in the t = 4 iterations are 4, 5, 3, and 3, respectively. The results of the mutation are four chains with lengths 8, 7, 10, and 9, respectively. The subroutine *Mutation* returns the offspring (1, 2, 3, 6, 9, 18, 36, 37) of length 7 instead of 10.

e	1	2	3	6	9	15	17	20	29	35	37
$r_2 = 4$	1	2	3	6	12	24	30	36	37		
$r_2 = 5$	1	2	3	6	9	18	36	37			
$r_2 = 3$	1	2	3	5	10	13	18	20	33	36	37
$r_2 = 3$	1	2	3	4	7	14	28	32	35	37	

Fig. 2. Four Mutated Chains for e=(1,2,3,6,9,15,17,20,29,35,37).

E. Survivor Selection

Survivor selection is the process of generating the next population from two population sets, where the first set is the current population and the second is the set of offspring that is generated from the mutation of the current population.

The method of selecting the survivors is done by combining the two sets, current and offspring, into a set of chains and then sorting the combined set based on the length of the chain in increasing order of length. After the sorting process has been completed, the first n smallest chains are selected based on length, such that the selected chains contain different elements. The two chains, E_i andE_j, are different if (1) the length of the two chains is different or (2) there exists at least one position, *pos*, such that the element at *pos* in the chain E_i is not equal to the element at *pos* in the chainE_j.

Our algorithm for survivor selection is different than that proposed in the previous work [16] in that our algorithm eliminates the step of calculating the fitness of a chain from q randomly selected chains. Also, our modified survivor selection algorithm involves deleting duplicated chains, i.e., two chains of the same length and containing the same elements are deleted. To do this, we use a different method to identify duplicated chains. The complete steps of the process of selecting the next generation are given in Subroutine SurvivorSel.

```
Subroutine SurvivorSel($E, O$)
Begin
    $Temp = E \cup O$
    Sort($Temp$)
    $Aux_1$= $Temp_1$
    $i = 1, k = 1$
    Repeat
        $i = i + 1$, flag=false, $j = i - 1$
        While ($j > 0$) and (flag=false) do
            If ($|Temp_i| \neq |Temp_j|$) then
                flag=true, $k = k + 1$
                $Aux_k = Temp_i$
            Else If (Equal($Temp_i, Temp_j$))
                flag=true
            $j = j - 1$
        If (flag=false) then
            $k = k + 1$
            $Aux_k = Temp_i$
    Until ($k \geq n$) or ($i \geq |Temp|$)
    Return $Aux$
End
```

F. Complete Proposed Algorithm

The proposed algorithm starts by generating an initial population, *E*, consisting of *n* ACs. Each AC is a valid chain, meaning that it satisfies the conditions for an AC. Then the algorithm repeats a sequence of steps consisting of *MaxNoIter* iterations, where *MaxNoIter* is one of the parameters used in the EA and represents the maximum number of iterations. The first step in the repetition loop is the application of the mutation operator on the population set, *E*. The output of this step is a set of mutated ACs of size *n*. The second step in the repetition loop is the application of the survivor selection subroutine on the combination of the current population and mutated sets. The output of this step is the next generation of the population of ACs of size *n*. The complete pseudocode of the proposed algorithm is shown in Algorithm Evolutionary Addition Chain, EAC.

```
Algorithm EAC (Evolutionary Addition Chain)
Input: a natural number $e \geq 2$.
Output: a short chain $(e_0, \dots, e_l)$.
Begin
    InitialParameters($MaxNoIter, n, t$)
    $E$ =InitialPopulation($e, n$)
    For $i = 1$ to MaxNoIter do
        $O$ =Mutation($e, E, n, t$)
        $E$= SurvivorSel ($E, O$)
    Return $E_1$
End
```

V. Performance Evaluation of the EAC

In this section, we study the performance of the proposed algorithm (EAC) in terms of the length of the generated AC. In order to verify performance, we compare the output of the EAC, i.e., the length of AC, with that of the following types of algorithm:

- An exact algorithm, denoted as ExA.
- Two non-exact deterministic algorithms, namely, the binary algorithm (BA) and the continued fraction algorithm (CFA).
- Two non-exact non-deterministic algorithms, one based on the GA and one based on the EA.

The following subsections describe the experimental setup, including the machine and software used in the coding of the algorithms, as well as the initialization parameters and the data required to execute the EA. Then, in the second, third, and fourth subsections we introduce and explain the results of the experimental study for the first, second, and third type of comparison, respectively.

A. Experimental Setup

Several parameters can be used in an experimental study to assess their effects on the performance of AC algorithms that are based on AI techniques:

- The first parameter, $\lambda(e)$, is the size of the natural number e, which is equal to the number of bits, b. In our experiment we used $b = 10, 11, 12, 13$, and 14 for all studied algorithms.
- The second parameter, $\varphi(b)$, is the set of all numbers of size (number of bits), b, which is equal to 2^b. This means that $\varphi(b) = \{2^b, 2^b + 1, 2^b + 2, \dots, 2^{b+1} - 1\}$. This parameter was used for all studied algorithms in our experiment.
- The third parameter, *n*, is the size of the population. In our study, *n* = 100, in line with previous works. This parameter was used for all studied EA.
- The fourth parameter, *MaxNoIter*, is the maximum number of generations used in the computation. In our

experiment, we set *MaxNoIter* as 300 for all studied EA.

- The fifth parameter, *t*, is the number of mutated sequences. In our study, *t* = 4, similar to previous works [16]. This parameter was used for all studied EA.

All algorithms were implemented using C++ language on a computer running a Windows operating system, which had a 2.4-GHz processor and a 32-GB memory.

For a comparison between two algorithms, A and B, we first determined the number of bits *b* and generated all natural numbers of *b* bits, $\varphi(b)$. Second, we applied the two algorithms on the number $e = 2^b$ and measured the length of the output of both algorithms. Third, we increased the value of e by one. Then we repeated the second and third steps until the last element, $e = 2^{b+1} - 1$, was reached.

In our comparisons of two algorithms, we measured the following criteria that are related to the length of the chain:

- $\#(A = B)$, which is the number of cases where the length of the AC generated by both algorithms is equal.
- $\#(A < B)$, which is the number of cases where the length of the AC generated by algorithm *A* is less than that generated by algorithm *B*.
- $\#(A > B)$, which is the number of cases where the length of the AC generated by algorithm *A* is greater than that generated by algorithm *B*.
- $MaxDiff(A, B)$, which is the maximum difference in the length between the lengths of the ACs generated by algorithm *A* compared to those generated by algorithm *B* for a fixed number of bits, *b*. Note that $MaxDiff(A, B)$ is not necessarily equal to $MaxDiff(B, A)$.

B. EAC and ExA

In this subsection, we compare the proposed algorithm, EAC, with the exact solution, ExA. The minimal length of ACs for *e* can be obtained from [37].

Table I displays the results of implementing ExA and EAC on different values of *b* as described in the experimental setup subsection. The results in the table lead to the following observations:

- For a fixed value of *b*, the output, i.e., the length of the AC, of both algorithms is the same in most of the cases, in that the percentage of $\#(EAC = ExA)$ is greater than 75%.
- The percentage of $\#(EAC > ExA)$ increases with an increase in the value of *b*. This means that with an increase in the value of *b*, the parameter *MaxNoIter* should be increased to obtain AC with short length near to the shortest length.
- When there is a difference in the lengths, the length of the AC generated by the EAC is near to the minimal length for the studied cases, because the maximum difference between the short and the shortest chains is 2.
- Even where the difference in the length of the AC generated by both algorithms is greater than 1, the percentage of such cases is small, in that it is less than 2%. For example, for *b* = 13 and 14, the percentage of cases that have a length greater than the minimal length of 2 is 0.5% and 1.9%, respectively.

C. EAC, BA and CFA

In this subsection, we compare the proposed algorithm, EAC, with two deterministic non-exact algorithms, BA and CFA that are based on the Fermat strategy [11].

Table II displays the results of comparing the performance of EAC and BA. These results lead to the following observations.

- The EAC performs better than the BA for all values of *b*, because the length of the AC generated by the EAC is less than or equal to that obtained by BA.
- The length of the AC generated by EAC is less than that obtained by BA in most of the studied cases with a percentage more than 85%.
- The maximum difference between the output of both algorithms increases with an increase in the number of bits, *b*. On the other hand, the percentage of the number of equal cases decreases with an increase in *b*. For example, the maximum difference between the two algorithms is 7, 8, and 9 for *b* = 12, 13, and 14, respectively.
- Most of the differences in the length of the AC generated by EAC and BA occur when *MaxDiff* = 1, 2, 3, and 4, as illustrated in Fig. 3.

TABLE I. COMPARISON OF EAC AND EXA

Criteria	***b***				
	10	**11**	**12**	**13**	**14**
$\#(\boldsymbol{EAC = ExA})$	1017 (99.3%)	1961 (95.7%)	3706 (90.5%)	6897 (84.2%)	14042 (76.5%)
$\#(\boldsymbol{EAC > ExA})$	7 (0.7%)	87 (4.3%)	390 (9.5%)	1295 (15.8%)	2342 (23.5%)
MaxDiff $(\boldsymbol{ExA, EAC})$	1	1	1	2 (0.5%)	2 (1.9%)

TABLE II. COMPARISON OF EAC AND BA

Criteria	***b***				
	10	**11**	**12**	**13**	**14**
$\#(\boldsymbol{EAC = BA})$	145 (14.2%)	207 (10.1%)	296 (7.2%)	455 (5.6%)	677 (4.1%)
$\#(\boldsymbol{EAC < BA})$	879 (85.8%)	1841 (89.9%)	3800 (92.8%)	7737 (94.4%)	15707 (95.9%)
$\#(\boldsymbol{EAC > BA})$	0	0	0	0	0
$\boldsymbol{MaxDiff(EAC, BA)}$	5	7	7	8	9
$\boldsymbol{MaxDiff(BA, EAC)}$	0	0	0	0	0

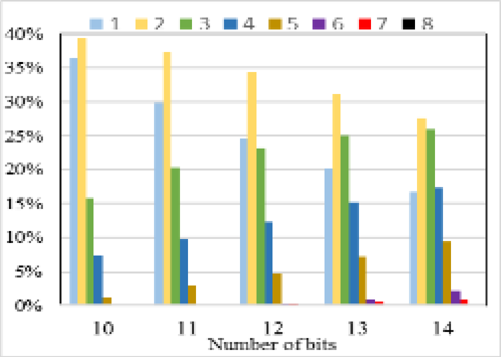


Fig. 3. Distribution of difference in Lengths for EAC< BA.

Table III displays the results of comparing EAC and CFA. These results lead to the following observations:

- The EAC shows better performance than the CFA for all values of *b* with percentage more than 25%. Also, the length of the AC generated by both algorithms is equal in more than 65% of all cases.
- In a few cases (less than 5%), the length of the AC generated by the CFA is less than that obtained by the EAC.
- The maximum difference between the two algorithms is limited to 2.
- When $\#(EAC < CFA)$, most cases occur when *MaxDiff* = 1, as illustrated in Fig. 4.

D. EAC and GA

In this subsection, we compare the proposed algorithm, EAC, with the GA. The details of the GA and its pseudocode can be found in [38].

Table IV displays the results of running the EAC and GA on different values of *b*. The results lead to the following observations:

- The EAC outperforms GA for all values of *b*, because the length of the AC generated by the EAC is less than or equal to that obtained by GA in 99% of cases.
- The percentage of $\#(EAC < GA)$ increases with an increase in the value of *b*. For example, the percentage of cases that have $\#(EAC < GA)$ for *b* = 11 and 12 are equal to 65.87% and 71.8%, respectively.
- When the GA performs better than the EAC, the maximum difference in the length of ACs is 1. On the other hand, when the EAC performs better than the GA, the maximum difference in the length of the AC increases with an increase in *b*.
- Most of the differences in the AC lengths generated by the EAC and the GA occur when $MaxDiff(EAC, GA)$ =1, 2 and 3, as illustrated in Fig. 5.

Remark: We compared EAC with the EA that is proposed by [16], we found that there is no significant difference between the output of both algorithms, i.e., minimal length of ACs. The main difference between them is in the running time, where our algorithm is faster than the EA in [16] with almost 25%. The reasons for reducing the running time come from (1) using the variable diff during generation the elements of chains, and (2) our proposed method for survivor selection.

TABLE III. COMPARISON OF EAC AND CFA

Criteria	***b***				
	10	**11**	**12**	**13**	**14**
$\#(\boldsymbol{EAC} = \boldsymbol{CFA})$	772 (75.4%)	1433 (70.0%)	2805 (68.5%)	5507 (67.2%)	10856 (66.3%)
$\#(\boldsymbol{EAC} < \boldsymbol{CFA})$	252 (24.6%)	610 (29.8%)	1262 (30.8%)	2536 (31.0%)	4934 (30.1%)
$\#(\boldsymbol{EAC} > \boldsymbol{CFA})$	0	5 (0.2%)	29 (0.7%)	149 (1.8%)	594 (3.6%)
$\boldsymbol{MaxDiff}$ $(\boldsymbol{EAC}, \boldsymbol{CFA})$	2	2	2	2	2
$\boldsymbol{MaxDiff}$ $(\boldsymbol{CFA}, \boldsymbol{EAC})$	0	1	1	1	2

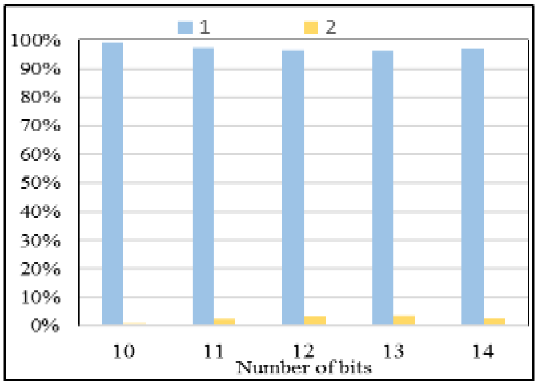


Fig. 4. Distribution of difference in Lengths for EAC< CFA.

TABLE IV. COMPARISON OF EAC AND GA

Criteria	***b***				
	10	**11**	**12**	**13**	**14**
$\#(\boldsymbol{EAC} = \boldsymbol{GA})$	435 (42.5%)	697 (34.0%)	1148 (28.0%)	1920 (23.4%)	2281 (13.9%)
$\#(\boldsymbol{EAC} < \boldsymbol{GA})$	589 (57.5%)	1349 (65.9%)	2941 (71.8%)	6255 (76.4%)	14103 (86.1%)
$\#(\boldsymbol{EAC} > \boldsymbol{GA})$	0	2 (0.1%)	7 (0.2%)	17 (0.2%)	0
$\boldsymbol{MaxDiff}(\boldsymbol{EAC}, \boldsymbol{GA})$	4	4	5	6	7
$\boldsymbol{MaxDiff}(\boldsymbol{GA}, \boldsymbol{EAC})$	0	1	1	1	0

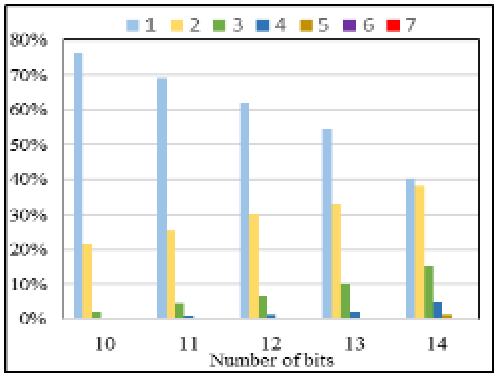


Fig. 5. Distribution of difference in Lengths for EAC< GA.

VI. Conclusion and Future Works

The AC can be used to decrease the number of multiplications in the encryption procedure of the RSA cryptosystem. In order to use the AC effectively, it is necessary to develop a method to find the shortest AC or a short AC. In this work, we discussed how we modified the EA to find a short AC in an efficient and simple way by focusing on the main components of EA such as representation, population, mutation and survivor selection. The proposed algorithm, EAC, was then implemented and compared with three types of algorithm (exact, non-exact deterministic, and non-exact non-deterministic) to assess its effectiveness. The experimental results indicated that the EAC showed good performance in comparison with the other types of algorithm when applied to natural numbers of 10, 11, 12, 13, and 14 bits.

In future work, we will extend our proposed algorithm to deal with large numbers of bits. We will also compare the performance of our algorithm with a wider range of deterministic and non-deterministic algorithms. Furthermore, we will consider running time as another criterion in the performance comparison.

Acknowledgment

This work has been funded by Scientific Research Deanship at the University of Ha'il – Saudi Arabia through project number RG-191309.

References

[1] H. Bahig, "Improved generation of minimal addition chains," Computing, vol. 78, pp. 161–172, 2006.

[2] H. Bahig, "On a generalization of addition chains: Addition–multiplication chains," Discrete mathematics, vol. 308, no. 4, pp. 611-616, 2008.

[3] H. Bahig, "Star reduction among minimal length addition chains," Computing, vol. 91, pp. 335–352, 2011.

[4] H. Bahig, "A fast optimal parallel algorithm for a short addition chain," J Supercomputing, vol. 74, no. 1, pp. 324-333, 2018.

[5] H. Bahig, and K. AbdElbari, "A fast GPU-based hybrid algorithm for addition chains," Cluster Computing, vol. 21, pp. 2001–2011, 2018.

[6] H. M. Bahig, A. Alghadhban, M. Mahdi, K. Alutaibi, and H. Bahig, "Speeding up the multiplication algorithm for large integers", Engineering, Technology & Applied Science Research, vol 10, no. 6, pp 6533-6541, 2020.

[7] H. Bahig, and H. Bahig, "A new strategy for generating shortest addition sequences," Computing, vol. 91, no. 3, 285-306, 2011.

[8] H. Bahig, H. Bahig, and K. Fathy, "Fast and scalable algorithm for product large data on multicore system," Concurrency and Computation: Practice and Experience, online published 2019, https://doi.org/10.1002/cpe.5259.

[9] H Bahig, and A Mahran, "Efficient generation of shortest addition-multiplication chains," Journal of the Egyptian Mathematical Society, vol. 26, no. 3, pp. 509-521, 2018.

[10] H. Bahig, and Y. Kotb, "An efficient multicore algorithm for minimal length addition chains," Computers, vol. 8, no. 1, pp. 1-23, 2019.

[11] F. Bergeron, J. Berstel, and S. Brlek, "Efficient computation of addition chains," J. de Theorie Nombres de Bordeaux, vol. 6, pp. 21-38, 1994.

[12] N. Clift, "Calculating optimal addition chains," Computing, vol. 91, pp. 265–284, 2011.

[13] N. Cruz-Cortés, F. Rodríguez-Henríquez, R. Juárez-Morales, C. Coello, "Finding optimal addition chains using a genetic algorithm approach," Lecture Notes in Computer Science, vol. 3801, pp. 208-215, 2005.

[14] N. Cruz-Cortes, F. Rodriguez-Henriquez and C. Coello, "An artificial immune system heuristic for generating short addition chains," IEEE Transactions on Evolutionary Computation, vol. 12, no. 1, pp. 1-24, 2008.

[15] Cruz-Cort´es, Nareli,c Rodr´ıguez-Henr´ıquez, Francisco, Ju´arez-Morales, Ra´ul and Carlos A. Coello Coello (2005). Finding optimal addition chains using a genetic algorithm approach. Y. Hao et al. (Eds.): CIS 2005, Part I, LNAI 3801, 208–215.

[16] S. Dominguez-Isidro, E. Mezura-Montes, and L. Osorio-Hernandez, "Evolutionary programming for the length minimization of addition chains," Engineering Applications of Artificial Intelligence, vol. 37, pp. 125–134, 2015.

[17] P. Downey, B. Leong, and R. Sethi, "Computing sequences with addition chains," SIAM Journal on Computing, vol. 10, no. 3, pp. 638-646, 1981.

[18] K. Fathy, H. Bahig, H., Bahig, and A. Ragb, "Binary addition chain on EREW PRAM," Lecture Notes of Computer Science, vol. 7017, pp. 321-330, 2011.

[19] K. Fathy, H. Bahig, and A. Ragab, "A fast parallel modular exponentiation algorithm," Arabian Journal for Science and Engineering, vol. 43, pp. 903–911, 2018.

[20] D. Gordon, "A survey of fast exponentiation methods," Journal of Algorithms, vol. 27, no. 1, pp. 129-146, 1998.

[21] R. Goundar, K. Shiota, and M. Toyonaga, "New strategy for doubling-free short addition-subtraction chain," Applied Mathematics & Information Sciences, vol. 2, no. 2, pp. 123–133, 2008.

[22] D. Knuth, "The Art of Computer Programming: Seminumerical Algorithms," vol. 2, 1973, Addison-Wesley.

[23] A. León-Javier, N. Cruz-Cortés, M. Moreno-Armendáriz, and S. Orantes-Jiménez, "Finding minimal addition chains with a particle swarm optimization algorithm," Lecture Notes in Computer Science, vol. 5845, pp. 680-691, 2009.

[24] A. Jayaram, and S. Deb, "A hybrid addition chaining based light weight security mechanism for enhancing quality of service in IoT," Wireless Personal Communications, vol. 113, pp. 1073–1095, 2020.

[25] K. Jrvinen, V. Dimitrov, and R. Azarderakhsh, "Generalization of addition chains and fast inversions in binary fields," IEEE Trans Computers, vol. 64, no. 9, pp. 2421–2432, 2015.

[26] A. Mullai, and K. Mani, "Enhancing the security in RSA and elliptic curve cryptography based on addition chain using simplified swarm optimization and particle swarm optimization for mobile devices," International Journal of Information Technology, online published 2020, https://doi.org/10.1007/s41870-019-00413-8.

[27] N. Nedjah, and L. de Macedo Mourelle, "Minimal addition chain for efficient modular exponentiation using genetic algorithms," Lecture Notes in Computer Science, vol. 2358, pp. 88-98, 2002.

[28] N. Nedjah, and L. de Macedo Mourelle, "Minimal addition-subtraction chains using genetic algorithms," Lecture Notes in Computer Science, vol. 2457, pp. 303-313, 2002.

[29] N. Nedjah, and L. de Macedo Mourelle, "Finding minimal addition chains using ant colony," Lecture Notes in Computer Science, vol. 3177, pp. 642-647 , 2004.

[30] N. Nedjah, and L. de Macedo Mourelle, “High-performance SoC-based implementation of modular exponentiation using evolutionary addition chains for efficient cryptography,” Applied Soft Computing, vol. 11, no. 7, pp. 4302-4311, 2011.

[31] L. Osorio-Hernandez, E. Mezura-Montes, N. Cruz-Cortes, and F. Rodriguez-Henriquez, “A genetic algorithm with repair and local search mechanisms able to find minimal length addition chains for small exponents,” 2009 IEEE Congress on Evolutionary Computation, Evolutionary Computation, Trondheim, Norway, 18-21 May 2009, pp. 1422–1429.

[32] S. Picek, C. Coello, Jakobovic, D. Jakobovic, and N. Mentens, “Finding short and implementation-friendly addition chains with evolutionary algorithms. J Heuristics vol. 24, pp. 457–481, 2018.

[33] S. Sanchez, J. Osorno, and E. Camarillo, “Simulated annealing meta-heuristic for addition chain optimization,” European Journal of Electrical and Computer Engineering, vol. 3, no. 6, pp. 1-4, 2019.

[34] E. Thurber, “Efficient generation of minimal length addition chains,” SIAM J Computing, vol. 28, pp. 1247–1263, 1999.

[35] J. Yang and C. Chang, “Efficient residue number system iterative modular multiplication algorithm for fast modular exponentiation,” IET Computers & Digital Techniques, vol. 2, no. 1, 1-5, 2008.

[36] Yen, S.-M, “Cryptanalysis of secure addition chain for SASC applications. Electronics Letters, vol. 31, no. 3, pp. 175–176, 1995.

[37] Shortest Addition Chain: http://wwwhomes.uni-bielefeld.de/achim/addition _chain.html

[38] Gentic Addition Chain Algorithm: https://github.com/josip-u/add-chain-solver.

A User-centered Design Approach to Near Field Communication-based Applications for Children

Mrim Alnfiai
Information Technology Department
College of Computer Sienese and Information
Taif University, Taif, Saudi Arabia

Abstract—**There is an abundance of technology targeting children in terms of education, entertainment, and health; however, little research has been conducted on the usability of Near Field Communication (NFC) to create an interactive, digital environment for children easily accessible on a mobile device. NFC technology is a component of Radio Frequency Identification (RFID) technology and is affordable, intuitive, and accessible. The following research evaluates existing NFC applications for children in terms of their ease of use, appropriateness, and areas in need of improvement. Recommendations are provided in visual design, audio enhancements, reward system, and privacy and security concerns. It is concluded that adopting NFC technology in all facets of life will positively benefit the most vulnerable population, children, but first progress toward a user-centered design for this group is required.**

Keywords—Near Field Communication (NFC); mobile device applications; NFC tags; RFID; early learning; K-12 students; preschool children; educational software; usability; user-centered design

I. Introduction

In modern society, the number of children using smartphone applications and new technologies is increasing. According to a report, children ages 5 to 9 years old spend about 28 minutes online daily, and this time continuously grows [1]. Another study found that 69% of children, across eight countries, use a smartphone [1]. Yet another study found that “almost 80 per cent of children who access the Internet from smartphones download or use mobile apps” [1]. Children are using a variety of smartphone applications for different purposes, like entertainment, learning, and taking care of their health. Selecting an appropriate program for a child based on his/her age and inclination is an important task, which requires parents and researchers to pay attention to an application’s key features. Therefore, developers need to design an application that suits the child's desires, builds his/her skills, and develops his/her confidence. On the contrary, if developers design an inappropriate program, this may negatively affect the child's skills and reduce the child's self-confidence.

Numerous research studies have shown that children do not yet possess the capabilities and knowledge of adults, so they are considered a different group with their characteristics [2], [3], [4]. Thus, it is preferable to study this group’s properties individually by taking a user-centered design approach. A variety of applications have been designed to provide services for children, to enhance the quality of their daily lives, and to provide adequate support to promote involvement in the activities designed by developers. However, while various applications have been developed for children, most of them do not encourage the user to explore and interact with their physical environment. This is despite a set of applications and systems already established in the form of NFC technology, which can be more immersive and user-friendly for children.

This paper serves as a review of existing NFC programs in terms of user-centered design for children. Recommendations are proposed to take into consideration the needs of this vulnerable population. The organization is as follows: an explanation of Near Field Communication technology, a review of existing NFC technology targeting children, analysis of the lack of user-centered design concepts in terms of NFC technology for children, a discussion of recommendations for areas of improvement, and concluding thoughts with a look at future research for NFC technology use in children’s applications.

II. Near Field Communication Technology

Most new smartphone devices are compatible with Near Field Communication (NFC) technology. NFC is a wireless technology that works at 13.56 MHz and is used for short-distance connections. It is a small chip used to store data and can be read by smartphone devices enabled with NFC technology. It supports two communication modes to enable two-way interactions, which are the Peer-To-Peer Mode and the Reader/Writer Mode. In the Peer-To-Peer Mode, a user can simply touch two smartphone devices together to establish the connection and share content between devices. In the Reader/Writer Mode, a user can touch an NFC tag with an NFC-compatible smartphone device to read the tag content. A smartphone-enabled with the NFC tag can read the content of a tag placed elsewhere from a 10 cm distance [5], [36].

NFC technology is a part of Radio Frequency Identification (RFID) technology; both of allow two-way communication. But RFID technology requires a special RFID reader, unlike NFC technology, which uses an NFC-enabled smartphone to read the chip content. NFC allows a short-range wireless information transfer between smartphone devices, thus the interaction between the devices is simply by touch. Therefore, it is a secure transfer, as it prevents other devices from reading the content of the NFC tag from a distance. Other security issues have been identified but are still being addressed by NFC developers.

The advantages of NFC technology are that it is simple, easy to use, and easy to learn [6], [7]. It is clear how to interact with the NFC, so it does not require a great amount of cognitive load by children [6]. Additionally, other researchers report that it only requires touching the target tag, which is an effortless process to communicate and interact with devices [7], [36]. Another advantage of NFC technology is that it is quick, as it requires only seconds to establish the connection and read the tag content.

III. Related Work

Researchers are currently employing NFC features to develop engaging tools for children. NFC technology has been extended to various fields to promote children's engagement in these areas: education, transportation, safety, entertainment, and health care fields [8], [9], [10], [11], [12], [37], [38], [39], [40], [41], [42], [43], [44].

A. Education

Several applications with the support of NFC technology have been proposed to enhance learning environments for children [8], [9], [10], [11], [12]. For example, some of these applications are used to support youth learning a second language, while others are used to track school attendance and transportation.

1) Learning a foreign language: Sánchez, Cortés, Riekki, and Oja (2011) present two interactive NFC-based applications to improve the experience of learning languages, which are Touch&Learn Languages and Touch&Learn Reading [10]. The primary purpose of both applications is to allow children to learn new vocabularies and learn how to read a foreign language. In the Touch&Learn Languages application, a user touches an NFC tag with their smartphone, and the application shows the related information that is linked to the NFC content. For example, it displays a word and reads it. By doing so, a child can learn a new word and its pronunciation using their physical environment as an interactive setting. In the Touch&Learn Reading application, the application displays a word and reads it aloud, and the child must touch the corresponding object that is tagged with an NFC tag. If they touch the correct object, the application presents the object image as confirmation.

Similarly, Lorusso et al. (2018) developed an application that reads the content of NFC tags. The tags are attached to small, touchable samples of objects like plastic animals and toys [11]. The purpose of the application is to enrich the vocabulary and conceptual networks for children with language impairments. The application was tested by kindergarten-age children, and the study results indicated this application is a valuable, engaging learning tool. It also speeds up the learning process, as it makes learning engaging and enjoyable.

Sánchez, Riekki, Rousu, and Pirttikangas (2008) propose another NFC-based interactive application, which is called Touch&Share [13]. The application uses NFC technology to store information about an object or an animal within a museum. This information stored can be audio files, videos, or text documents to enhance learning. In this application, a user touches a tag that is attached to an item in the museum, and the application provides additional information about the touched item. This application provides an interactive, learning environment within the museum. The application was tested in the local Zoological Museum, and the NFC tags were attached to stuffed animals so that children could learn more about their habitat and life patterns. The study results show that using NFC technology provides an immersive experience in which children can learn.

2) Attendance supervision: Ervasti, Isomursu, and Kinnula (2009) developed an attendance supervision system with the support of NFC technology. In the system, each student was given an NFC card that has their ID (the child's name) [14]. When students arrive at school, they touch an NFC reader, like a smartphone device, that is located at the class entrance with their NFC cards. At that time, the NFC reader records the student ID, the direction (in or out), and arrival time. Students repeat the same procedure when they leave at the end of the school day. The system allows parents to track their children's' attendance without coming to school by sending an instant message to them. It also improves and secures children's transitions from home to school and vice versa. In addition, it enables students to be independent when traveling between home and school, as well as enhances communication between school and home. Another benefit of the attendance supervision system is that it helps teachers by minimizing responsibilities like marking absences. The critical benefit the school gets from the system is that it discourages absenteeism by informing teachers, administrators, and parents of a child's absence in real-time, allowing for immediate intervention. The system was tested in a school by teachers, Grade 1 students, and parents. They report that it has made a marked difference in attendance as well as received a positive adoption rate and provided improvements in the school's daily routine. However, some technical concerns have been reported about the system, which are privacy and security issues such as protecting student information including name, location, and arrival and departure time.

3) Transportation: Rengaraj and Bijlani (2016) presented a child security and monitoring model with NFC/RFID technology and a smartphone device to track child activity and ensure safety [15]. The system is a web-portal and Android-based mobile application that uses NFC/RFID technology, a smart ID card, and a Child Safety device (CSD). The CSD contains RFID and Global System for Mobile Communications (GSM), or Internet-enabled Wi-Fi or Ethernet cables to communicate with a server. The model provides instant communication between parents, children, and teachers to ensure security in the school, on the bus, and at home by providing a report about the children's activities. The solution was evaluated, and the study outcome shows an acceptance of the use of the system for child security at both parenting and school management levels.

Recently, Bai, Fu, and Yang (2018) proposed an NFC-based system on a smartphone device for children's safety when they are picked up from school. The proposed system allows teachers to monitor children's movements by identifying and ensuring the pick-up person is rightfully in charge of the child to be picked-up [18]. Teachers scan the authorized person's NFC tag and the child's NFC tag, then the system compares the two tags to ensure they match. If they match, the system sends a message to the child's parents, which includes the name and ID of the individual who has already picked up their child. On the other hand, if the comparison does not match, then the system will automatically communicate with the parents to inform them to correct the authorized individual's information. The system was not evaluated. However, there is a need to add warning functions if the child is absent or they are not at the correct pick-up place.

4) Pervasive learning system: Ivanov proposed a pervasive learning service with the support of NFC technology that describes the objects that are tagged to assist children in recognizing objects within their physical environment [8]. Children can use the system just by touching the objects around them that are tagged by NFC tags to learn their names and characteristics. The system syncs with the Google App Engine cloud, which stores audio recording information about each object to provide verbal information to children. The service can be used in different learning scenarios such as colors, letters, and numbers, shape recognition, learning foreign languages, and many others. The system was evaluated with 10 children ages 3-8 years-old and the results showed that the system was easy to use, and the interface was accessible for children. Findings also reported that the system can be used as an interactive teaching tool, which allows children to learn in their own environment.

B. Health Care

Many NFC applications have been developed to automate healthcare services, including health promotion, patients' records, and patients' identification.

1) Nutrition guidance (mHealth): Vazquez-Briseno et al. (2013) introduced a smartphone application with the support of NFC technology and QR codes. It offers nutrition guidance to promote healthy diets for children and prevent obesity [19]. The main page of the application presents three buttons, which are the NFC tag scan button that reads the content of the tag attached to food; the second button, which provides a set of food groups; and the last button, which presents and explains the Eat Well Plate. After scanning the tag, the application displays stars as a motivation method if the child ate correctly. Parents are also informed about their child's food selections. Another important feature in the application is, that if a child eats more than the maximum calorie intake, the child and their parents will receive a warning notification. It has been tested in an elementary school environment, and it was an effective method of minimizing health-related problems.

In 2020, Sutjiredjeki et al. presented a medical measurement application with the support of NFC technology and the IoT (Internet of Things) [16]. The application measures five health parameters, which are height, weight, body temperature, blood oxygen (SpO2), and heart rate. The application enhances the examination process and speeds it up, as well as allows better integration and management of the children's health records. The main reason for using NFC technology is to automatically transmit the collected data from different sensors to the NFC-based smartphone devices and the electronic key management system (e-KMS) tags, which in this case is electronic medical records for Indonesian children. Using NFC technology in this application eliminates manual recording and digitizes patient records. The result shows that the proposed system can be used to measure several health parameters quickly, and it minimizes the time spent to examine patients.

C. Entertainment

Riekki, Sasin, and Pirttikangas (2008) developed a smartphone game with the support of NFC technology; it is called Touch&Run [17]. The goal of the application is to improve the learning experience and to encourage children to do physical exercise in teams. The game board is divided into a 4*4 grid, and each grid has an NFC tag, which includes the position of that square on the board and the state of that square in the game. A player and their team must occupy the grid as fast as possible by scanning its NFC tag to encourage communication between team members. The team players touch their phones with each other for a few seconds, which allows the phones to transmit between each other. The Touch&Run game allows children to learn teamwork skills and cooperation between team members.

In 2019, Chen et al. proposed an interactive gaming application called intelligent SOMA (iSOMA) [20]. It attaches an NFC tag on each side of the iSOMA block to determine and detect the current position and the relative position of other blocks. This way, the application can assist users in checking whether they correctly built the blocks and to quickly assemble iSOMA structures. Using NFC technology in the iSOMA game helps in learning the process of how to build the cube quickly, and it overcomes the limitations of the traditional SOMA cube.

Imanara and Horie (2017) developed a home appliance controller with the support of NFC technology [21]. In their proposal, a child touches an NFC tag attached to a plastic toy that represents a house appliance. In return, the command transfers through a Wi-Fi router and enables the corresponding physical household appliance to operate. The evaluation study indicated this game has the chance to improve children's cognitive development by discovering the relationship between the plastic toy and the actual household appliance.

IV. RESEARCH CHALLENGES AND RECOMMENDATIONS

Several NFC applications and evaluation studies have been conducted to better understand applications intended to assist children in completing varying tasks. Nevertheless, there is a lack of study that discusses and analyzes the usability of NFC applications for children. I analyzed the existing NFC applications for children. From a user-centered design perspective, most NFC technologies targeting children are in need improvement to meet children's needs, wants, and desires. As a result, recommendations for improvements to overcome these design challenges are presented. Kraleva (2017) and Kraleva, Kralev and Kostadinova (2016) have highlighted some important factors, which should also be taken into consideration when taking a user-centered design approach to developing an NFC-based application for children [22], [23]. These factors include specifying the targeted age group, following simplified design principles, adding audio effects, ensuring that the application is easy to use and rewards success, and taking into account privacy and security roles.

A. *Defining the Target Age Group*

Most, but not all, of the proposed NFC applications stipulate the proposed user's age range and cognitive capabilities. Ibarra (2011) reported that children in between the ages of 2 and 4 years old are not able to concentrate on a digital game, unlike those in the age range between 5 and 6 years old, who can focus on the game's goal and can enjoy the concept behind it [24].

Table I shows which developers of NFC applications determined the target age group in their proposal application. Leaving the target group's age undefined negatively affects the acceptance of the applications among children. As each group has a particular ability and desires, developers need to specify for whom their application is built.

B. *Simplified Design*

Most user-centered design studies focused on children encourage developers to simplify the design of an application's interface; it should have limited and large buttons that are labeled clearly to enable children to easily recognize the interface items [25]. Some researchers recommend using different colors to get children's attention and avoid using menus, to prevent confusion. Most of the applications that use NFC technology to target young users follow a simple design, and they apply a set of colors to get children's attention.

C. *Music and Sounds*

Salmon (2009) found that using sound and music promotes children's thinking and improves their learning development [26]. She also recommended that teachers and parents use music to enhance children's learning and cognitive development. However, none of the NFC applications for children evaluated in this research provides musical accompaniment, even though it was found by several studies that using proper sounds for kids' applications makes them better perceived by children [27].

The Nutrition Guidance application developed by Vazquez-Briseno et al. (2013) provides textual alerts and notifications, but this is unsuitable for children [19]. Children who fall in the age range between 2 and 4 years old are not able to read text; thus, text alerts are not appropriate for them. Using multimedia like sound, videos, and pictures to deliver the message is a better method to notify children and keep them interested [19]. After thorough review, it can be concluded that developers failed to take this into consideration as 42% of the applications studied do not provide visual and aural feedback (as shown in Table I).

D. *Easy Tasks*

Developers should simplify tasks to avoid making children bored, which could cause them to avoid using the application the next time. Most NFC applications studied (92%) are simple, as they only require children to touch the NFC tags with their NFC-enabled mobile device, without any need to input text; meaning, these applications are overwhelmingly suitable for young children, as they do not require typing any content to interact with the application interface. The application not utilizing easy tasks is the Touch&Run.

TABLE I. NFC APPLICATIONS FOR CHILDREN BASED ON DESIGN PRINCIPLE

Applications	**Features**								
	Defining the target age group	***Simplified design***	***Music and sounds***	***Easy tasks***	***Prizes***	***Privacy policy***	***Testing***	***Usability***	***Accuracy***
Touch&Learn Reading (2011)	Undefined	√	√	√	√	×	√	√	×
Lorusso et al. [11]	preschool 4-6 years old	√	√	√	√	×	√	√	×
Touch&Share (2011)	children 9-13 years old	√	√	√	×	×	√	√	×
Attendance supervision system	First grade, Special-need Class	√	×	√	×	×	√	√	×
Monitoring system [15]	Undefined	√	×	√	×	×	√	√	×
Pervasive learning system [8]	Aged3 -8 years old	√	√	√	×	×	√	√	×
Children Picking system [18]	Undefined	√	×	√	×	×	×	√	×
Nutrition Guidance [19]	Children, adults	√	×	√	√	×	√	√	×
A medical application [16]	Aged 1-12 months	√	×	√	×	×	√	√	×
Touch&Run [17]	Preschool	√	√	×	×	×	√	√	×
iSOMA [20]	Undefined	√	√	√	×	×	√	√	×
Home appliances' controller [21]	Children 3+	√	√	√	×	√	×	√	×

E. Prizes

A recent study highlighted that rewarding children improves their accuracy and encourages them to practice efficient engagement [28]. Prizes also encourage children to keep learning and motivate them to continue doing the required tasks. To keep children's attention and make them concentrate more, developers should reward them after completing any task, such as with a pleasant melody, asterisks, or any engaging image [29], [35]. For example, the Touch&Learn application presents stars on the application interface when a child completes a task correctly [10]. In the NFC-based applications reviewed, only 25% of the programs used a prize or rewards system.

F. Privacy Policy and Security

The most important element that developers consider in any smartphone application is security. Researchers mention the importance of preserving the data of children and protecting those under 13 years of age by a privacy policy, to protect them from the electronic world and to keep them safe [30]. The involvement of NFC technology with smartphone technology introduces new potential security issues, such as eavesdropping [31], data corruption [32], and a man-in-the-middle attack [33].

The security concern is critical when it comes to financial transactions, but the NFC technology is considered secure in most other regards because its communication range is very limited. However, children’s information, including name, age, education, and location, should be encrypted to protect children’s private information and prevent this information from being leaked to unauthorized persons or hackers. To minimize security issues for NFC applications, some developers ask for a login to ensure that only authorized children and their parents can access the application.

As previously references, some NFC applications are built to track children, like the school attendance or transportation program. Tracking children’s movements by parents and teachers is a valuable method for child safety, but it can also be a risk when the captured data is not encrypted. Unfortunately, 92% of the NFC applications reviewed do not address these privacy and security issues and do not encrypt collected data. Thus, these applications need to be refined based on a security perspective and for children’s safety.

G. Testing

Testing is an important part of developing user-centered software. Since the proposed NFC technology is developed for a special group of users, namely children, the applications must be stable and of the highest quality. It is important to involve the end-users in the design and evaluation phase of the application. Most of the NFC applications for children reviewed (83%) have been tested by a few children and their parents, and the results indicate the children’s acceptance of the NFC technology and its simplicity to learn. However, a large number of NFC applications for children have not yet been verified and evaluated in the real-world environment. After testing the functionality of the application, longitudinal usability studies are necessary to assess the acceptance of the proposed system by children, family members, and teachers. In addition, the effects of the proposed system on the child's education and skill development must be evaluated. Furthermore, the consequences and disadvantages of the system on the child's behavior must be evaluated for further development, such as the impact of these applications on children’s emotions and motivations. Finally, the most critical aspect that needs to be tested in children’s applications is that related to security and privacy issues, as there is a serious need to provide a secure and safe technical world for children.

Other studies highlight some important criteria that developers should take into consideration, which are:

H. Usability

A new interaction method has been introduced once developers began using NFC technology. Several studies indicated that any new interaction method would cause unconsidered usability concerns [34]. Thus, conducting a set of user studies is critical to expose the usability challenges and meet children’s capabilities. It is important to identify the application’s usability problems from a user-centered design approach. As shown in Table I, all of the NFC applications were tested for usability. They were found to be usable, as most of the application solved a problem and achieved the goal, and they were easy to learn (and easy to remember) how to use. In addition, the NFC applications that were discussed in related works offer appropriate functions so that children can do what they want to do, and they can simply complete their tasks.

I. Accuracy

Accuracy is a very important criterion that encourages children to use the developed application several times. However, none of the NFC applications’ accuracy has been measured as seen in Table I. Building accurate applications allow children to use the service in a safe environment with straightforward steps. The error rate can be minimized by building a prototype that addresses most of the proposed application functions and having the target group evaluate it to clearly understand the target group’s ability. Developers must take into account the accuracy of the information about the child, especially for the NFC applications that are designed for the healthcare field. Then, the accurate information like location, supports the proposed system to provide the wanted service.

V. Discussion

The review of existing NFC-based applications targeted to children shows the technology to be a valuable tool for children in various fields including education, health care, and entertainment. It also finds that most of the NFC applications are cost-effective and require little time or investment to accomplish tasks, as they enable the quick capture of the tag content and transfer it just by touch. NFC technology in education plays an important role, as it is not time-consuming to exchange data.

The findings also reveal that NFC technology is suitable for users with low or no technical background, particularly among children. The NFC applications evaluated seek to contribute to the development of children. As these applications provide the opportunity for self-study at children's homes these applications enable an accessible interactive learning environment. Considering these findings, the integration of NFC technology and smartphone applications into early childhood education is thought to be of great importance.

However, employing NFC technology to support children’s learning and remaining motivated is considered an emerging practice, as currently there is no widespread use of NFC-based applications. Developers need to consider the impact of other factors when designing NFC applications for children to achieve the successful functionality of their programs. These user-centered considerations include targeted age, knowledge and skills level, simplicity and ease of use, NFC tag locations, social influences, accuracy, and design consistency [10].

One of the advantages of NFC technology is that it can be integrated easily with smartphone applications, and it is usable and easy to use and learn; thus, children can use NFC-based applications with no need for training or support. Another advantage is that all of the NFC-based applications evaluated use a simple design, so children become familiar with the technology and learn how to use it in a short time. However, these NFC-based applications do not meet all inherent children’s needs as 42% of those applications reviewed do not consist of music or sound to attract a user’s attention. In addition, 75% of the applications studied need to provide some engagement or rewarding methods, like showing stars or providing clapping sounds after achieving a task [29].

Another limitation is that the accuracy of the NFC-based applications mentioned has not been evaluated, which is important to minimize the rate of error and minimize their effect, especially in applications that track children’s locations. It is very important to improve these tracking and monitoring applications by addressing privacy and security issues, to protect children’s information, and to provide a secure digital space for children. There is also a need to implement security protocols when using NFC technology, to secure the data exchange of children’s information.

In the future, there is a need to improve NFC-based applications targeting children be implementing a user-centered design approach.

A. Strengths and Limitations

NFC technology has been used for various applications and it has proved successful in most cases. NFC is used to improve user interaction, as it can afford children the ability to identify the objects around them while encouraging them to do physical activity and share knowledge quickly with others. NFC-based applications has positive qualities that enables it to be effective in many fields and to become increasingly popular due to its high efficiency, low cost, convenience, interaction technique, high reliability, and ability to integrate with smartphones [1], [8], [21], [26], [31].

Another advantage of this technology is that the shape of NFC tags can be customized. Moreover, NFC tags can be used anywhere, because they are waterproof. In addition, the time needed to read the NFC tag content is short. Thus, most of the applications that use NCF technology are simple and efficient. However, as previously noted, it is vulnerable to attack by hackers. Most of the discussed NFC applications for children (92%) do not address a privacy policy or a particular method to secure children’s and parents’ information.

VI. Conclusion

By exploring NFC applications, children can use NFC technology to leverage and enhance learning, communication, entertainment, and healthcare. The adoption of NFC technology in schools, hospitals, and the home allows children to positively interact with their surroundings and actively learn while doing physical activities.

The programs developed using NFC technology were studied based on the effective design principles for children and the design criteria for mobile programs in general. Implementing this user-centered design approach, results indicate that some NFC applications have many strengths, which are taking into account the capabilities of the end-users and their age; ease of use; and using some media, such as pictures and sound, to reflect the program’s reaction to the child’s behavior. However, some of these applications missed important criteria, including adding sound or rewarding children after completing the required tasks. In addition, developers of smartphone applications that use NFC technology need to take into account the privacy of children and increase the protection of personal information and program content, as well address the urgent need to increase the accuracy of data presented by NFC applications. Thus, there is a serious need to improve applications that use NFC technology targeting children, as it is considered an enabler technology that creates usable and easy-to-learn interactive environments for children in the fields of learning, communication, entertainment, and healthcare. Otherwise, it is a missed opportunity.

References

[1] Gutnick, A. Robb, M. L. Takeuchi, J. Kotler, L. Bernstein, and M. Levine. Always connected: The new digital media habits of young children. The Joan Ganz Cooney Center at Sesame Workshop, 2011.

[2] Alnfiai, M., Sampalli, S., & MacKay, B. (2016). VirtualEyez: Developing NFC technology to enable the visually impaired to shop independently. International Journal of Electrical and Computer Systems (IJECS),1-13. DOI: 10.11159/ijecs.2016.001.

[3] Butler, J. R., & Nelson, N. L. (2020). Children overclaim more knowledge than adults do, but for different reasons. Journal of Experimental Child Psychology, 201.https://doi-org.sdl.idm.oclc.org/10.1016/j.jecp.2020.104969.

[4] Du, Y., Clark, J. E., Valentini, N. C., Kim, M. J., & Whitall, J. (n.d.). Children and adults both learn motor sequences quickly, but do so differently. Frontiers in Psychology, 8(FEB). https://doi-org.sdl.idm.oclc.org/10.3389/fpsyg.2017.00158.

[5] Nikitin, P. V. Rao K. V. S. and Lazar, S. (2007). An Overview of Near Field UHF RFID. IEEE International Conference on RFID, Grapevine, TX, 2007, pp. 167-174, doi: 10.1109/RFID.2007.346165.

[6] Rukzio, E., Leichtenstern, K., Callaghan, V., Schmidt, A., Holleis, P., & Chin, J. (2006). An experimental comparison of physical mobile interaction techniques: Touching, pointing and scanning. Proceedings of the 8th International Conference on Ubiquitous Computing (UbiComp 2006: Ubiquitous Computing). Springer. doi: 10.1007/11853565_6.

[7] Välkkynen, P., Niemelä, M., & Tuomisto, T. (2006). Evaluating touching and pointing with a mobile terminal for physical browsing. Proceedings of the 4th Nordic Conference on Human-Computer Interaction (NordiCHI ’06). Association for Computing Machinery,

New York, NY, USA, 28–37. DOI:https://doi.org/10.1145/1182475.1182479.

[8] Vanov, R. (2013). NFC-based pervasive learning service for children. In Proceedings of the 14th International Conference on Computer Systems and Technologies (CompSysTech '13) (pp. 329-336). Association for Computing Machinery, New York, NY, United States. DOI:https://doi-org.sdl.idm.oclc.org/10.1145/2516775.2516804.

[9] Emilia Biffi, Peter Taddeo,Maria Luisa Lorusso, and Gianluigi Reni. 2014. NFC-based application with educational purposes. In Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare, 370-372. Doi: 10.4108/icst.pervasivehealth.2014.255350.

[10] Sánchez, I., Cortés, M., Riekki, J., & Oja, M. (2011). NFC-based interactive learning environments for children. In Proceedings of the 10th International Conference on Interaction Design and Children (IDC '11). Association for Computing Machinery, New York, NY, USA, 205–208. DOI:https://doi.org/10.1145/1999030.1999062.

[11] Lorusso ML, Biffi E, Molteni M, Reni G. (2018). Exploring the learnability and usability of a near field communication-based application for semantic enrichment in children with language disorders. Assist Technol. 2018;30(1):39-50. doi: 10.1080/10400435.2016.1253046. Epub 2017 Feb 13. PMID: 28632018.

[12] Kim, K., Jeong, S., Kim, W., Jeon, Y., Kim, K., & Hong, J. (2017). Design of small mobile robot remotely controlled by an android operating system via bluetooth and NFC communication. 2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI). Jeju, South Korea. doi: 10.1109/URAI.2017.7992864.

[13] Sánchez, I., Riekki, J., Rousu, J., & Pirttikangas, S. (2008).Touch & Share: RFID based ubiquitous file containers. In Proceedings of the 7th International Conference on Mobile and Ubiquitous Multimedia (MUM '08). Association for Computing Machinery, New York, NY, USA, 57–63. DOI:https://doi.org/10.1145/1543137.1543148.

[14] Ervasti, M., Isomursu, M., & Kinnula, M. (2009). Experiences from NFC supported school attendance supervision for children. 2009 Third International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies. Sliema, Malta. doi: 10.1109/UBICOMM.2009.9.

[15] Vinoth Rengaraj and Kamal Bijlani. (2016). A Study and Implementation of Smart ID Card with M-Learning and Child security, 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), July 2016, pp. 305-311.

[16] Sutjiredjeki, Ediana & Basjaruddin, N & Fajrin, Diki & Noor, F. (2020). Development of NFC and IoT-enabled measurement devices for improving health care delivery of Indonesian children. Journal of Physics: Conference Series. 1450. 012072. 10.1088/1742-6596/1450/1/012072.

[17] Riekki, J., Sasin, S., & Pirttikangas, S. (2008). Touchnrun: An RFID-based distributed board game motivating to move. Poster Proc. Persuasive Technology (Persuasive 2008), University of Oulu, Department of Information Processing Science, Series A, Research Papers, A42: 34-40.

[18] Bai, Y., Fu, C., & Yang, J. (2018). Using NFC tags and smartphones to design a reliable mechanism to pick a child up from school. 2018 IEEE International Conference on Consumer Electronics (ICCE).Las Vegas, NV, United States. doi: 10.1109/ICCE.2018.8326205.

[19] VAZQUEZ BRISENO, Mabel et al. mHealth Platform and Architectures to Provide Nutritional Guidance to Children. International Journal of Interactive Mobile Technologies (iJIM), [S.l.], v. 7, n. 4, p. pp. 15-20, oct. 2013. ISSN 1865-7923. doi:http://dx.doi.org/10.3991/ijim.v7i4.3083.

[20] Chen, C. et al. (2019). Intelligent SOMA interactive gaming system. 2019 IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin). Berlin, Germany. doi: 10.1109/ICCE-Berlin47944.2019.8966163.

[21] Imanara, S., & Horie, R. (2017). Implementations of dollhouse-smart house interface: Towards an intuitive household appliance control system for age 3+. 2017 IEEE 6th Global Conference on Consumer Electronics (GCCE). Nagoya, Japan. doi: 10.1109/GCCE.2017.8229321.

[22] Kraleva, R. (2017). Designing an interface for a mobile application based on children's opinion. International Journal of Interactive Mobile Technologies, 11(1), 53–70. https://doi-org.sdl.idm.oclc.org/10.3991/ijim.v11i1.609.

[23] Kraleva, R., Kralev, V., & Kostadinova, D. (2016). A conceptual design of mobile learning applications for preschool children.ArXiv, abs/1606.05753.

[24] Ibarra, K. (2011). Designing apps for kids. UX Magazine. Available at https://uxmag.com/articles/designing-apps-for-kids.

[25] Gelman, D. L. (2014). Design for kids. Digital products for playing and learning. Brooklyn, New York : Rosenfeld Media.

[26] Salmon, Angela. (2010). Using music to promote children's thinking and enhance their literacy development. Early Child Development and Care. 180. 937-945. 10.1080/03004430802550755.

[27] Tikkanen, R., & Iivari, N. (2011). The Role of Music in the Design Process with Children. Human- Computer Interaction -- INTERACT 2011: 13th IFIP TC 13 International Conference, Lisbon, Portugal. https://doi.org/10.1007/978-3-642-23765-2_21.

[28] Jin, X., Auyeung, B., & Chevalier, N. (2020). External rewards and positive stimuli promote different cognitive control engagement strategies in children. Developmental Cognitive Neuroscience, 44.

[29] Farrow, C., Belcher, E., Coulthard, H., Thomas, J. M., Lumsden, J., Hakobyan, L., & Haycraft, E. (2019). Using repeated visual exposure, rewards and modelling in a mobile application to increase vegetable acceptance in children. Appetite, 141. https://doi-org.sdl.idm.oclc.org/10.1016/j.appet.2019.104327.

[30] iUbend. (2016). A duide to COPPA and mobile apps. Retrieved from http://www.iubenda.com/blog/guide-coppa-mobile-apps/.

[31] Chattha, N. A. NFC — Vulnerabilities and defense. (2014). Conference on Information Assurance and Cyber Security (CIACS), Rawalpindi, 2014, pp. 35-38, doi: 10.1109/CIACS.2014.6861328.

[32] Nearfieldcommunication.org.(2020). Security concerns with NFC technology. Retrieved from http://www.nearfieldcommunication.org/nfc-security.html.

[33] Haselsteiner, E., & Breitfuß, K. (2006). Security in Near Field Communication (NFC) strengths and weaknesses. Retrieved from http://events.iaik.tugraz.at/RFIDSec06/Program/papers/002%20-%20Security%20in%20NFC.pdf.

[34] McNamara, N., & Kirakowski, J. (2006). Functionality, usability, and user experience: Three areas of concern. Interactions, 13, 26-28. 10.1145/1167948.1167972.

[35] Glatt, S. J. (2020). Initial responsiveness to reward attainment and psychopathology in children and adults: An RDoC study. Psychiatry Research, 289. https://doi-org.sdl.idm.oclc.org/10.1016/j.psychres.2020.113021.

[36] Coskun, V., Ozdenizci, B., & Ok, K. (2013). A survey on near field communication (NFC) technology. Wireless Personal Communations, 71, 259–294.

[37] Alnfiai, M., Sampalli, S., MacKay, B. (2016). VirtualEyez: Developing NFC Technology to Enable the Visually Impaired to Shop Independently. Proceedings of the 2nd International Conference on Computer and Information Science and Technology (CIST'16). Ottawa, Canada – May 11 – 12, 2016, Recognized as the Best Paper in the Conference.

[38] Alnfiai, M., Sampalli, S., MacKay, B. (2016). VirtualEyez: Developing NFC Technology to Enable the Visually Impaired to Shop Independently. International Journal of Electrical and Computer Systems (IJECS),1-13. DOI: 10.11159/ijecs.2016.001.

[39] Alnfiai, M., Sampalli, S. (2017). BrailleEnter: A Touch Screen Braille Text Entry Method for the Blind. The 8th International Conference on Ambient Systems, Networks and Technologies (ANT 2017), Procedia, Computer Science.

[40] Alnfiai, M., Sampalli, S. (2017). Social and Communication Apps for the Deaf and Hearing Impaired. International Conference on Computer and Applications (ICCA'17), Dubai, Computer Science.

[41] Alnfiai, M., Sampalli, S. (2017). An Evaluation of BrailleEnter keyboard: An Input Method Based on Braille Patterns on Touchscreen

Devices. International Conference on Computer and Applications (ICCA'17), Dubai, Computer Science.

[42] Alnfiai, M., Sampalli, S. (2018). BraillePassword: Accessible Web Authentication Technique on Touchscreen Devices. Journal of Ambient Intelligence and Humanized Computing. Springer.

[43] Garrido, P. C., Miraz, G. M., Ruiz, I. L., & Gomez-Nieto, M. A. (2011). Use of NFC-based pervasive games for encouraging learning and student motivation. In Near Field Communication (NFC), Third International Workshop on Near Field Communication, Hagenberg, 2011, pp. 32-37, doi: 10.1109/NFC.2011.13.

[44] Alan, U. D., & Birant, D. (2018). Server-Based Intelligent Public Transportation System with NFC. in IEEE Intelligent Transportation Systems Magazine, vol. 10, no. 1, pp. 30-46, Spring 2018, doi: 10.1109/MITS.2017.2776102.

Enhancement of Natural Language to SQL Query Conversion using Machine Learning Techniques

Akshar Prasad[1], Sourabh S Badhya[2], Yashwanth YS[3], Shetty Rohan[4], Shobha G[5], Deepamala N[6]
Department of Computer Science and Engineering
RV College of Engineering
Bengaluru, India

Abstract—**In the age of information explosion, there is a huge data that is stored in the form of database and accessed using various querying languages. The major challenges faced by a user accessing this data is to learn the querying language and understand the various syntax associated with it. Query given in the form of Natural Language helps any naïve user to access database without learning the query languages. The current process of conversion of Natural Language to SQL Query using a rule-based algorithm is riddled with challenges -- identification of partial or implied data values and identification of descriptive values being the predominant ones. This paper discusses the use of a synchronous combination of a hybrid Machine Learning model, Elastic Search and WordNet to overcome the above-mentioned challenges. An embedding layer followed by a Long Short-Term Memory model is used to identify partial or implied data values, while Elastic Search has been used to identify descriptive data values (values which have lengthy data values and may contain descriptions). This architecture enables conversion systems to achieve robustness and high accuracies, by extracting meta data from the natural language query. The system gives an accuracy of 91.7% when tested on the IMDb database and 94.0% accuracy when tested on Company Sales database.**

Keywords—Machine learning; natural language to SQL query; long short-term memory; embedding layer; elastic search; hybrid architecture

I. Introduction

Availability of data and its analytics have revolutionized our life in all aspects. One of the popular methods of storing and accessing data is using Structured Query Language (SQL). SQL is a domain-specific programming language designed to store and access data in relational databases. It requires professional skills to use it. With the demand for data increasing exponentially, a simpler querying method which requires lesser or no learning time is a necessity. Hence, significant efforts are on to make Natural Language an interface between humans and the data stored in computers. Querying database using Natural Language makes data access simpler and affordable by all users.

Currently majority of the conversion systems that convert Natural Language to SQL query employ rule-based algorithms [1], [2]. One of the main challenges of this method is to identify implied or partial data values in the Natural Language. Another frequent failure case of such algorithms is the inability to capture the lengthy data values, often the attributes which contain descriptions and hence referred to as 'descriptive values' in this paper. The proposed system aims to resolve these shortcomings and increase the accuracy and robustness of the Natural Language to SQL query conversion systems.

To understand the challenge of identifying implied/partial data values in a domain specific database, consider the following example Natural Language query given to a sales database: 'Get the price of *product* red scooter'. This can be understood by a rule-based algorithm that 'red scooter' is a product and has to be searched in the corresponding 'product' attribute. However, the query 'Get the price of red scooter' requires the system to understand that scooter is a data value of 'product' attribute. Another interesting query is, 'Get the price of red two-wheeler' which requires the system to be intelligent and robust to understand that the user is implying the data-value 'scooter' with the use of the term 'two-wheeler'.

The proposed system uses an embedding layer followed by an LSTM model to pick up n-grams which are similar to determine a data value with a confidence greater than a pre-set threshold. The system has also been equipped with WordNet to find hyponyms of attributes to pick up implied values, in cases where the schema vocabulary is not expressive enough to train the embedding layer well.

Descriptive columns have been dealt with separately in this architecture. For example, let a database maintained by a pet adoption centre, have an entry with 'Animal Name' as 'Baxter' with 'Description' as 'He is a highly active and enthusiastic six-month old dog. He is black in colour and loves to chase vehicles'. A Natural Language query, 'Name the six-month dog which is fun loving and dark in colour' would be highly challenging to convert to the right SQL query, without special care being taken to understand the semantics. The proposed system uses Elastic Search to identify such 'descriptive values.

In all the existing systems, either an unintelligent rule-based system is adopted or the machine learning models are burdened with the entire task of conversion. The proposed system, which adapts a few concepts implemented in [10] and [11], although essentially a rule-based algorithm, uses machine learning models, WordNet and Elastic Search to enhance the conversion by overcoming the challenges faced by rule-based systems. Hence the system is more robust and the accuracy of conversion increases by an approximated 11-16%.

The paper describes the architecture and the techniques used by the proposed system and tests the performance of the system on the IMDb database [3]. To test the performance on 'descriptive data values', the system is tested on Company Sales database. It is to be noted that when tested on the same database used to test the SQLizer [4], the architecture's accuracy is 13.8% higher, considering SQLizer's 'Top 1' results.

II. Related Work

Conversion of Natural Language to SQL query was first explored in 1992, by Nikolaus Ott et al [5] where Natural Language inputs were mapped to an augmented SQL language. Different approaches to such conversions, some of which involve machine learning are discussed further.

S. Javubar *et al* has used standard natural language techniques such as morphological analysis, semantic analysis, mapping tables for retrieval of reports from social web data [1]. Xiaojun Xu *et al* in their paper 'SQLNet: Generating Structured Queries from Natural Language Without Reinforcement Learning' attempt the conversion by filling in the slots of a standard SQL template with the data values present in the sentence using a CNN model. The performance is being considered on two different parameters, Query-Match with accuracy of 65.5% and Execution Accuracy of 71.5% [6].

Victor Zhong *et.al* in their paper 'Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning' use separate rules i.e. Neural Network models for each SQL clause. The different clauses considered are as follows

1) Aggregation clause – The tokens in the sentence are first mapped to its scalar attention score and these are normalized to obtain a distribution of input encodings. Sum of all the input encodings is taken and a multi-layer perceptron is used to convert the sum to a score corresponding to the aggregation (α). Softmax function is used to normalize the scores.

2) Select clause – Encoding of the column name with a LSTM and a multi-layer perceptron is applied over the column representations, conditioned on the input representation, to compute a score for each column.

3) Where clause – Reinforcement learning is applied to learn a policy to optimize execution results of expected correctness.

Finally, a mixed objective function is applied to combine the result of all the three clauses. The performance is considered on two different parameters, Query-Match Accuracy is 48.3% and Execution Accuracy of 59.4% [7].

Geordani et al. use concept of structured kernels e.g. Sequence and Tree Kernels which is referred to as structures that are created to classify the words present in the input query into appropriate tags. In this paper, a detailed comparison on different kernels were considered and appropriate combination were made to generate better results. Accuracies were considered for both datasets - Geo dataset (75.9%) and RestQueries dataset (84.7%) [8]. P. Utama *et al* in their paper 'An End-to-end Neural Natural Language Interface for Databases' use a novel concept of Neural Query Translation. An automatically generated dataset is fed into a Recurrent Neural Network and is used at runtime to convert natural language into SQL query. An Interactive auto-completion system increases translation accuracy since users enter less ambiguous sentences as input. Accuracies were considered for both datasets - Geo dataset – (48.6%) and Patients dataset (75.93%) [9].

F. Li *et al.* in their paper 'Constructing an interactive natural language interface for relational databases' explore the construction of parse trees and query trees to individually map words into their corresponding SQL tags. The system is commonly referred to as NaLIR (Natural Language Interface for Relational databases). An accuracy of 57.14% was achieved for the MAS dataset [2].

III. Proposed System

The proposed system extracts the dataset from the input relational database in CSV format. It must be noted that the system is not specific to a particular schema, but the models need to be re-trained if a new database is to be introduced. Three separate components work synchronously to extract maximum latent information from the dataset, which can either be used to enrich the natural language or be stored to use during conversion.

The system architecture has been described in Fig. 1. Each of the components described here is explained in detail in the following sections.

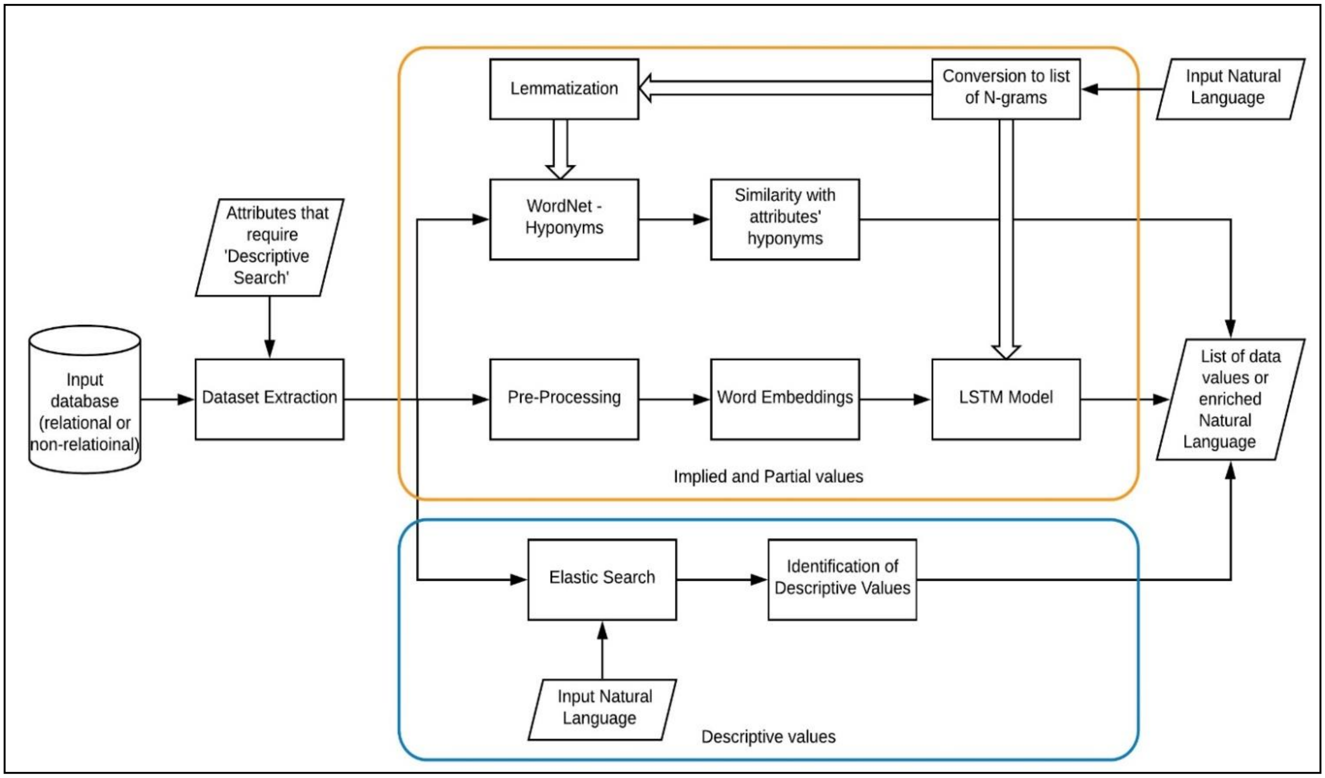

Lemmatization
Conversion to list of N-grams
Input Natural Language
WordNet - Hyponyms
Similarity with attributes' hyponyms
Attributes that require 'Descriptive Search'
Input database (relational or non-relatioinal)
Dataset Extraction
Pre-Processing
Word Embeddings
LSTM Model
List of data values or enriched Natural Language
Implied and Partial values
Elastic Search
Identification of Descriptive Values
Input Natural Language
Descriptive values

Fig. 1. System Architecture.

IV. Dataset Extraction

A major challenge faced while integrating machine learning into the architecture is to find a suitable dataset. The dataset is extracted from the input relational database. Apache Common CSV Library [12] has been used to extract the dataset in the form of CSV files. Along with the relational database, markers for attributes which contain descriptive values' (Ex: Experience, Description, etc.) is to be provided as input. Note that the system architecture is not schema specific and the ML model must be re-trained for a new database.

V. Partial and Implied Values

This section describes how the system identifies and extracts the partial and implied data values. This meta data can be used to either enrich the natural language input or be used directly by the rule-based conversion algorithm.

A. Pre-Processing Techniques

1) Reconstruction of dataset: The dataset extracted in the earlier step is a flat file containing the data values for all the attributes. A temporary feature vector is formed by obtaining the unique values of each attribute.

The data value and its attribute name separated by a placeholder value is used as the target value for each sample. This ensures that an extensive search to determine the attribute of the identified partial/implied value is avoided later.

2) Splitting, Tokenization and One-hot encoding: It was observed that the patterns in data values are best captured when the data value is split into words and each word is split into n-grams (value of n is determined by length of word). Intuitively, the syllables of each word of a data value are used as the features.

Each unique split (syllable) in the dataset is designated an integer (tokenized) as the model accepts only integer values. (The samples now consist of variable number of integers.) A label encoder followed by a one hot encoder is used to convert the target vector to an integer binary matrix.

3) Random sampling: To make the model robust and to expose the word embedding layer to a more varied vocabulary, a random sampling technique for augmentation is employed.

Iterating through each sample, a variable number of tokens/integers (the range of which is calculated based on the length of the sample) is randomly chosen for a fixed number of times (based on the number of samples) and appended to the dataset along with its target value (which is a binary vector). It is to be noted that random sampling does not affect the ordering of the remaining tokens.

4) Padding and Truncation: To ensure that all samples are of uniform length, the samples are padded with trailing zeros. The samples are then truncated to the 75th quartile of lengths to ensure that most features are not the padded zeros. Repeating the features of data values which have very few tokens increased the accuracy of detection of such attributes/data values.

The pre-processing steps have been summarized in Fig. 2.

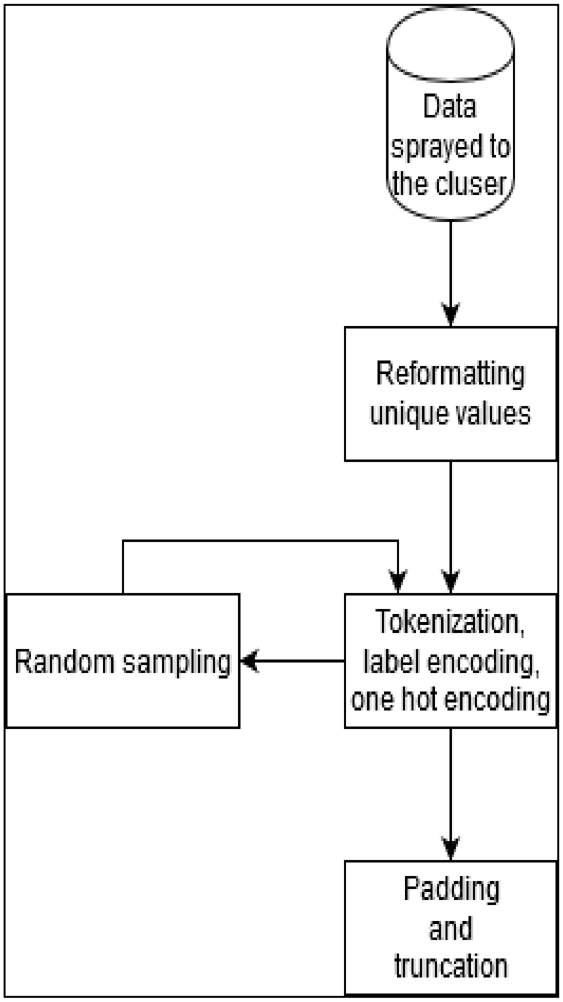

Data sprayed to the cluser
Reformatting unique values
Tokenization, label encoding, one hot encoding
Random sampling
Padding and truncation

Fig. 2. Pre-Processing Summary.

B. Embedding Layer

Word embeddings are representations of words in an n-dimensional vector space. This representation has helped bridge the gap between machine's and human brains' understanding of the language. An embedding layer maps each word in the given corpus to a dense vector which represent the projection of the word into a specified dimensional space. The vectors for the words are learnt based on its surrounding words in the given corpus. As a result, the vectors of words with similar meaning will be 'close to each other'. For example, the Euclidean distance between the vectors representing 'school' 'college' will be much lesser than between vectors representing 'school' and 'dog'.

Hence this approach succeeds in capturing the context of a word or sentence. This can be observed in Fig. 3, where the verb tense relationships and country-capital relationships have been recognized by the trained embedding layer.

C. Long Short-Term Memory (LSTM)

LSTM is a variant of Recurrent Neural Networks (RNN). RNNs which succeeded in creating a perception of storage or memory, often failed due to exploding and vanishing gradients. This failure case was overcome by LSTM with the use of a memory cell. The memory cell contains the current memory of the node which can be written into, read and erased just like a computer memory. (Note: This memory is analogous.)

The current timestep and the previous timestep's output is fed as an input to the LSTM node as seen in Fig. 4. The node contains a memory cell and four simple one-layer neural networks. While one neural network generates the new memory, two other neural networks control the significance given to the old memory and the new memory, and the other neural network generates the output from the new memory. Note that a 'tanh' activation function is used for memory generation, while a sigmoid activation is used to determine the significance.

Since LSTM specializes in processing series of samples where the temporal locality carries great significance, it has been used in this architecture to understand a series of dense vectors and classify them into the correct data value.

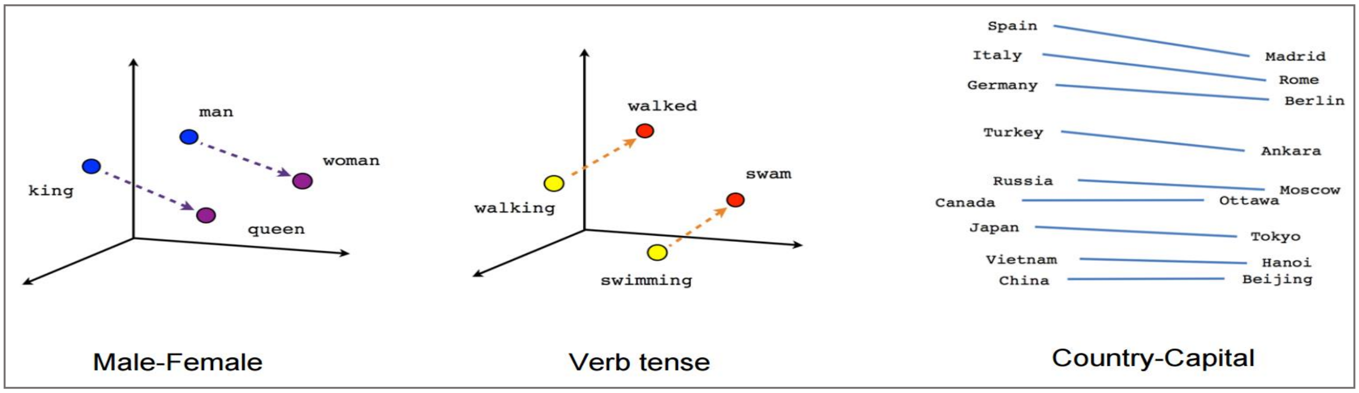


Fig. 3. Embedding Layer [13].

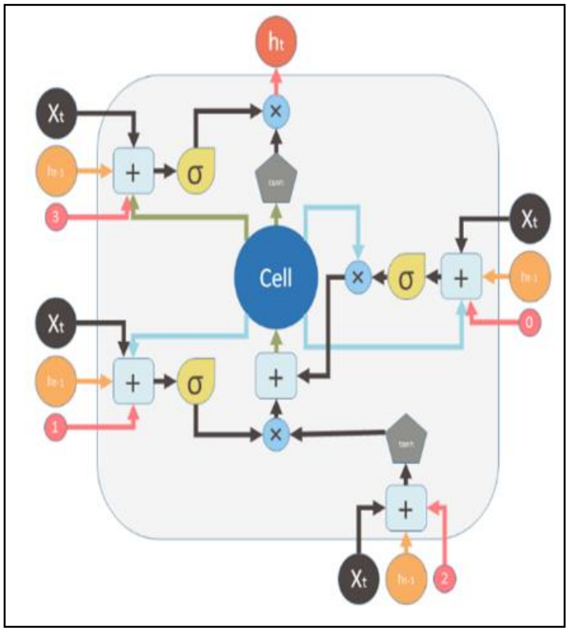


Fig. 4. Long Short-Term Memory [14].

The number of nodes used and the number of epochs are relatively small because of the following reasons:

1) Due to the nature of the dataset, there are high chances of over-fitting.

2) Speed of training the model must be high, as the model is re-trained every time a new schema is introduced.

D. Classification of Inputs

The input Natural Language query is tokenized and split into different sequences. Sequences of 1 word (1-gram) up to sequences of n words (n-gram, where n is determined by the number of tokens) is considered for prediction. The largest sequences and its classification are considered (i.e., sub-sequences are ignored).

The final, high confidence classifications given by the LSTM model can be used in multiple ways, couple of them are outlined below:

1) Enrich the natural language query: Replace the classified data value with the incomplete/implied data value (largest n-gram) and precede the attribute before the data value.

2) Store the data values and attribute names: Store the data value- attribute pair using a HashMap or any list and use it as required by the conversion algorithm.

VI. Descriptive Values

This section describes the elastic search approach taken for identifying and mapping the descriptive data values. These values are lengthy and tend to have high degrees of incompleteness and implications.

A modified version of the embedding layer and LSTM model approach was tried for such values as well. The approach's accuracy exceeds the elastic search approach only if the data values' vocabulary is expressive enough to develop a suitable vector space while training the embedding layer. As this is rare, the Elastic Search approach has been explored in this paper.

A. Architecture

Descriptive sentences are those sentences that provide broader context to the query. The extracted CSV file is used to create an index in Elastic Search. Elastic Search's Bulk API [15] provides the necessary functions that can create and store large data simultaneously. The 'helpers' module [16] in Python which is one of the ways to access this API is being used in this case to create and store the data from the .csv file. The benefit of using Bulk API is that it indexes chunks of data at once rather than one after the other. The input query is made to pass through the following to get the required information:

1) Analyzers: Analyzers are used to process the input query to a desired format. The analyzers are a class of in-built functions that bring the data to a standard format. Elastic Search has in-built analyzers such as Stop analyzer and Standard analyzer. Custom analyzers can be built to comply with the needs of the application. The system in this case makes use of 'Stop' and 'English Language' Analyzer [17], [18].

a) Stop Analyzer: This analyzer splits the text into individual words and removes all the pre-defined stop-words which are defined for English. Stop Analyzer also lower cases the remaining words after the removal of stop words. Elastic Search also supports Stop Analyzers in multiple languages.

Ex: *Input*: Get the doctors with master's degree.

Analyzer: Get doctors master's degree

It is to be noted that 'with' and 'the' tokens are discarded.

b) English Language Analyzer: This analyzer converts the words of the input query to its root word. This is used when all the words in the input query have different tenses. This analyzer brings them to their basic form. This makes the processing of the remaining words easier.

Ex: *Input*: Show all products which are red bikes.

Analyzer: Show all product which road bike.

It is to be noted that 'bikes' has changed to 'bike' and 'products' has changed to 'product'.

2) Searching through multiple attributes: Multiple columns can be searched in Elastic Search. This can be done by using the "multi_match" keyword [19]. The JSON request body which is to be sent to Elastic Search server can be written as follows:

```
{ "query":
        { "multi_match":
                { "query": input query,
                 "fields":[list of descriptive column names];
                }
        }
}
```

Format: Query for multiple attributes

The input query mentioned in the JSON request is the output obtained after the input query has been processed from the analyzers. “fields” represents the set of column names through which descriptive information needs to be searched for.

Ex: *Input*: Show all items which are road bikes.

Analyzer: show all product which road bike.

The request query is as follows:

```
{ “query”:
          { “multi_match”:
                        { “query”:“ list all item which road bike”,
                         “fields”: [‘UserDescription’,
‘ItemDescription’];
                        }
          }
}
```

Example: The input query is being searched across:

‘UserDescription’ and ‘ItemDescription’ fields of the index.

Once the request is sent, the response is received in the form of JSON. The JSON body contains useful information such as the “source” field which describes the appropriate description along with the field to which it's being matched.

In addition to this, there also exists a field called “_score” which indicates how relevant the description is to the matched description. The “_score” is calculated by Elastic Search by making use of a practical scoring function [20]. The practical scoring function is described in Elastics Search as follows:

$$score(q,d) : \\ = coord(q,d) \\ - queryNorm(q).?(tf(t\ in\ id).idf(t)^2.(t.getBoost()).norm(t,d))$$

Where, **t** refers to the term which is given in the input by the user, **q** is the cosine similarity between the vectors.

coord(q,d) represents a score dependent on how many query terms a given document contains.

tf(t in d) represents the term frequency of the term, t in the current document d.

queryNorm(q) normalizing factor to alter queries’ scores to make them comparable.

idf(t) stands for Inverse Document Frequency of t.

t.getBoost() is a search time boost of the term in the current query q.

norm(t,d) summarizes a few boost and length factors (indexing time).

Hence, this scoring function is dependent on the number of times a word in the input query appears in the documents that are present in the index. Intuitively, the relevance score of a descriptive value increases as the frequency of occurrence in documents increases.

Furthermore, a reasonable threshold for contextual relevance scoring must be set. Values above this limit are only to be considered for fieldname-value pair extraction.

3) Searching through multiple attributes: After the generation of pairs of fieldnames and values, a statement which corresponds to SQL can be written as follows-

WHERE fieldname$_1$ = value$_1$ AND fieldname$_2$ = value$_2$ AND.… fieldname$_N$ = value$_N$;

Later, the following statement is appended to existing SQL output which does not handle descriptive or contextual columns. If there is no suitable pair then the statement is not generated and there is no effect on the output.

A summary of these steps is depicted in Fig. 5.

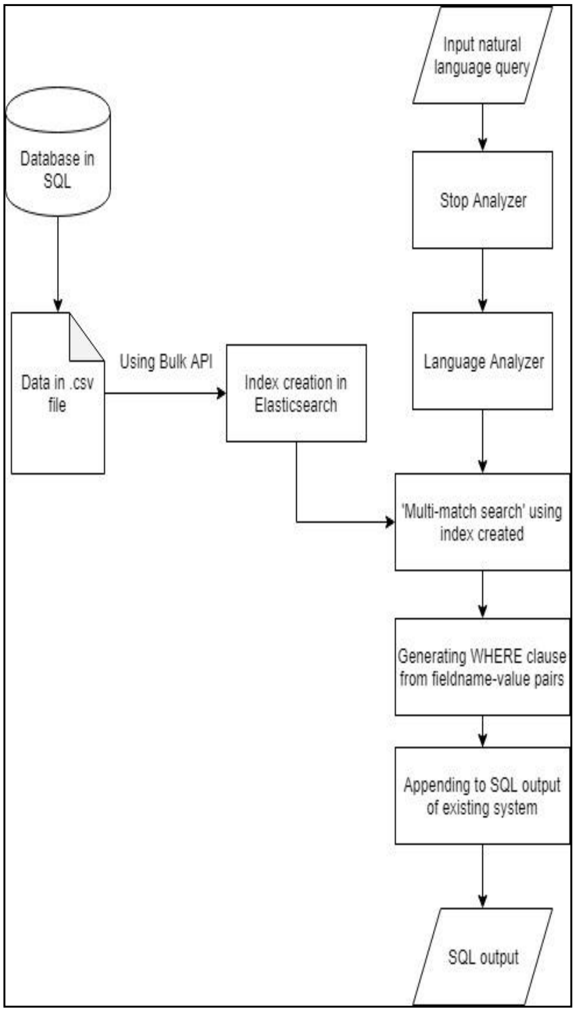
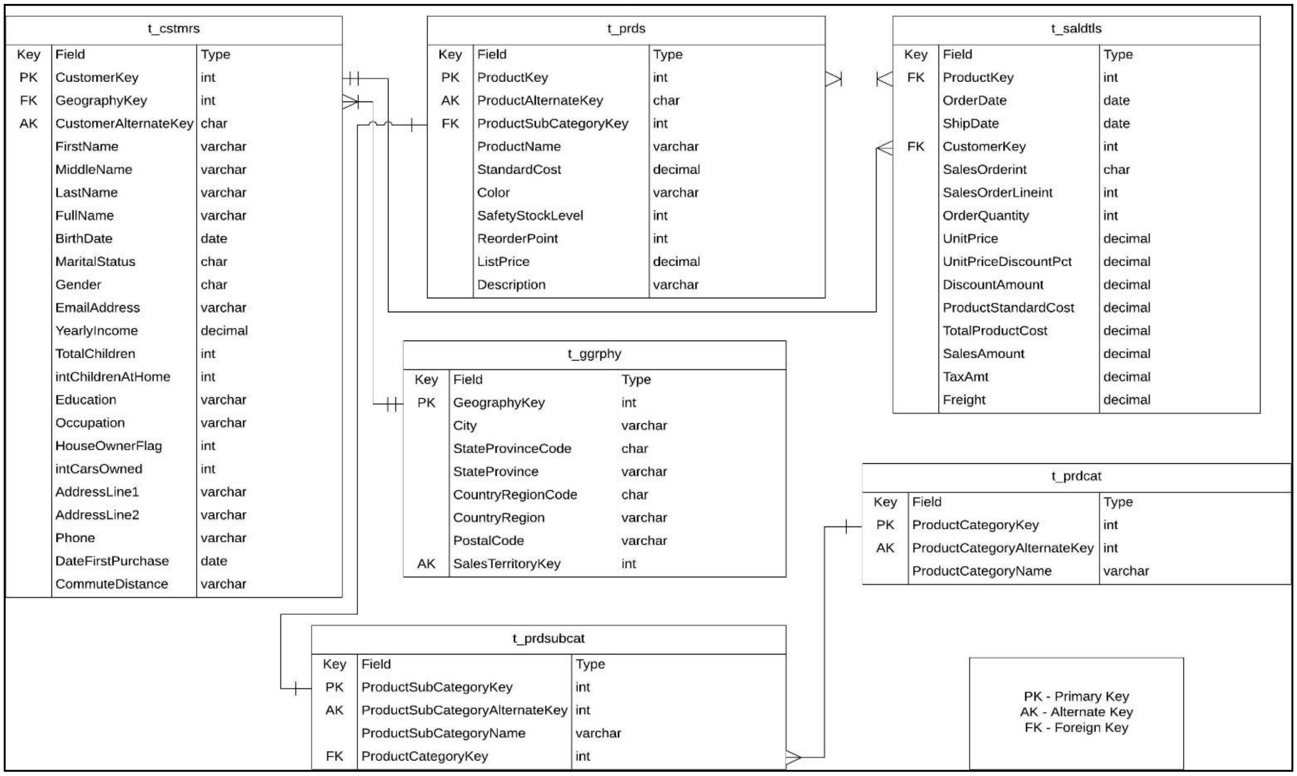

Input natural language query
Database in SQL
Stop Analyzer
Data in .csv file
Using Bulk API
Index creation in Elasticsearch
Language Analyzer
'Multi-match search' using index created
Generating WHERE clause from fieldname-value pairs
Appending to SQL output of existing system
SQL output

Fig. 5. System Architecture – Descriptive Values.

VII. Results

A. Company Sales Database

The system is tested on the Company Sales database, whose schema is depicted in Fig. 6. The database has six tables which contain the data of products, customers and its sales.



t_cstmrs

Key	Field	Type
PK	CustomerKey	int
FK	GeographyKey	int
AK	CustomerAlternateKey	char
	FirstName	varchar
	MiddleName	varchar
	LastName	varchar
	FullName	varchar
	BirthDate	date
	MaritalStatus	char
	Gender	char
	EmailAddress	varchar
	YearlyIncome	decimal
	TotalChildren	int
	intChildrenAtHome	int
	Education	varchar
	Occupation	varchar
	HouseOwnerFlag	int
	intCarsOwned	int
	AddressLine1	varchar
	AddressLine2	varchar
	Phone	varchar
	DateFirstPurchase	date
	CommuteDistance	varchar

t_prds

Key	Field	Type
PK	ProductKey	int
AK	ProductAlternateKey	char
FK	ProductSubCategoryKey	int
	ProductName	varchar
	StandardCost	decimal
	Color	varchar
	SafetyStockLevel	int
	ReorderPoint	int
	ListPrice	decimal
	Description	varchar

t_saldtls

Key	Field	Type
FK	ProductKey	int
	OrderDate	date
	ShipDate	date
FK	CustomerKey	int
	SalesOrderint	char
	SalesOrderLineint	int
	OrderQuantity	int
	UnitPrice	decimal
	UnitPriceDiscountPct	decimal
	DiscountAmount	decimal
	ProductStandardCost	decimal
	TotalProductCost	decimal
	SalesAmount	decimal
	TaxAmt	decimal
	Freight	decimal

t_ggrphy

Key	Field	Type
PK	GeographyKey	int
	City	varchar
	StateProvinceCode	char
	StateProvince	varchar
	CountryRegionCode	char
	CountryRegion	varchar
	PostalCode	varchar
AK	SalesTerritoryKey	int

t_prdcat

Key	Field	Type
PK	ProductCategoryKey	int
AK	ProductCategoryAlternateKey	int
	ProductCategoryName	varchar

t_prdsubcat

Key	Field	Type
PK	ProductSubCategoryKey	int
AK	ProductSubCategoryAlternateKey	int
	ProductSubCategoryName	varchar
FK	ProductCategoryKey	int

Fig. 6. ER Diagram of Company Sales Database.

1) Experimental Setup: The tables 't_prds' contains product data, 't_cstmrs' contains customer data, 't_saldtls' contains sales data, 't_prdcat' contains product-category data, 't_prdsubcat' contains product-subcategory data and 't_ggrphy' contains the geographical location of the customers.

The attribute 'Description' in 't_prds' table contain 'descriptive values' which describes each product. This column is marked while giving the database as input to the system.

2) Partial and Implied Values – Results: Of the 50 natural language queries that were designed to test this feature, 48 queries correctly identified the partial/implied data values and 47 queries were mapped to the correct SQL query.

The queries tested the ability of the system to generate queries with clauses such as WHERE, BETWEEN, HAVING, ORDER BY, GROUP BY, COUNT, SUM, MAX, MIN (aggregate functions).

Table I shows some of the correctly mapped natural language queries. The data values which were partial or implied have been highlighted in bold and the identified attributes have been underlined.

3) Descriptive Values – Results: The system was tested for 50 descriptive inputs and all the inputs mapped to appropriate descriptions and 49 of them formed to the correct SQL query. Results that were used for testing are shown in Table II. The descriptive data values have been highlighted in bold.

E. IMDB Database

The system is tested on the IMDb database whose schema is shown in Fig. 7. The database has tables that contain data of movies and its genres, actors, and directors.

1) Experimental Setup: The tables names are self-explanatory and the database overall contains movies', actors' and directors' data and are linked through 'roles' and 'movies_directors' tables.

The genres of movies and directors are also contained in the database.

2) Partial and Implied Values – Results: Out of the 48 natural language queries that were designed to test this feature, 44 queries correctly identified the partial/implied data values and mapped to the correct SQL query.

The queries again tested the ability of the system to generate queries with clauses such as WHERE, BETWEEN, HAVING, ORDER BY, GROUP BY, COUNT, SUM, MAX, MIN (aggregate functions).

Table III contains some of the correctly mapped natural language queries. The data values which were partial or implied have been highlighted in bold and the identified attributes have been underlined.

Note, the IMDb database has no scope for testing descriptive values. Company Sales database has been used for this purpose.

TABLE I. COMPANY SALES DATABASE: PARTIAL / IMPLIED VALUES

Input Natural Language Query	**Enriched Natural Language Query**	**Output SQL Query**
show all unmarried customers who are men	show all ***single*** Gender '***male***' customers	SELECT * FROM t_cstmrs WHERE LOWER(MaritalStatus) = 'single' AND LOWER(Gender) = 'male'
Names of customers who have graduated and from germany or france	FullName Names of customers who have Education '***graduate degree***' and from CountryRegion 'germany' or CountryRegion 'france'	SELECT t_cstmrs.FullName FROM t_cstmrs INNER JOIN t_ggrphy ON t_ggrphy.GeographyKey = t_cstmrs.GeographyKey WHERE LOWER (t_ggrphy.CountryRegion) = 'germany' OR LOWER (t_ggrphy.CountryRegion) = 'france') AND (LOWER(t_cstmrs.Education) = 'graduate degree')
get the price of red or dark helmet	get the price of Color 'red' or Color '***black***' ProductSubCategoryName 'helmet'	SELECT ListPrice , Color FROM t_prdsubcat INNER JOIN t_prds ON t_prdsubcat.ProductSubCategoryKey = t_prds.ProductSubCategoryKey WHERE LOWER(Color) = 'red' OR LOWER(Color) = 'black'
how much does tire tube cost	how much does ProductName '***road tire tube***' cost	SELECT ListPrice , ProductName FROM t_prds WHERE LOWER(ProductName) = 'road tire tube'
get the orders from new south wales australia	get the orders from StateProvince 'new south wales' CountryRegion 'australia'	SELECT t_saldtls.OrderQuantity, t_ggrphy.CountryRegion, t_ t_cstmrs.FullName , t_ggrphy.StateProvince FROM t_ggrphy INNER JOIN t_cstmrs ON t_cstmrs.GeographyKey = t_ggrphy.GeographyKey INNER JOIN t_saldtls ON t_cstmrs.CustomerKey = t_saldtls.CustomerKey WHERE LOWER(t_cstmrs.StateProvince) = 'new south wales' AND LOWER(t_ggrphy.CountryRegion) = 'australia'
show subtotal of orders for helmet	show subtotal of orders for ProductSubCategoryName '***helmet***'	SELECT SUM(t_saldtls.SalesOrderint) FROM t_prds INNER JOIN t_saldtls ON t_prds.ProductKey = t_saldtls.ProductKey WHERE LOWER(t_prds.ProductName) = 'helmet'
Names of customers in germany who purchased bicycle but not tires	FullName Names of customers in CountryRegion 'germany' who purchased ProductCategoryName '***bikes***' but not ProductSubCategoryName '***road tire tubes***'	SELECT t_cstms.FullName FROM t_cstmrs INNER JOIN t_ggrphy ON t_ggrphy.GeographyKey = t_cstmrs.GeographyKey INNER JOIN t_saldtls ON t_saldtls.CustomerKey = t_cstmrs.CustomerKey INNER JOIN t_prds ON t_prds.ProductKey = t_cstmrs.ProductKey INNER JOIN t_prdsubcat ON t_prdsubcat.ProductSubCategoryKey = t_prds.ProductSubCategoryKey WHERE LOWER(t_prdsubcat.ProductSubCategoryName) != 'road tire tubes' AND LOWER(t_prdsubcat.ProductSubCategoryName) = 'bikes' AND LOWER (t_ggrphy.CountryRegion) = 'germany'

TABLE II. COMPANY SALES DATABASE: DESCRIPTIVE VALUES

Input Natural Languge Query	**Output SQL Query**
list price of black products which is responsive, maneuverable, low budget and can go off-road.	SELECT t_prds.ListPrice FROM t_prds WHERE t_prds.Description = ***'This low budget bike gives a high performance.. It is maneuverable and responsive. Highly suitable to go off-road'*** AND (LOWER(t_prds.Color) = 'black'
Name the Road Series all round bike item for road and racing.	SELECT t_prds.ProductName FROM t_prdcat INNER JOIN t_prdsubcat ON t_prdcat.ProductCategoryKey = t_prdsubcat.ProductCategoryKey INNER JOIN t_prds ON t_prdsubcat.ProductSubCategoryKey = t_prds.ProductSubCategoryKey WHERE t_prds.Description = ***'Has specifications similar to our Road series bikes, but the frame size has been made to suit a woman. All round bike, suitable for road and racing'*** AND LOWER(t_prdcat.ProductCategoryName) = 'racing' AND LOWER(t_prdsubcat.ProductSubCategoryName) = 'bike')
Get the customers who have bought a bike which is comfortable, provides durable frame and high power	SELECT t_cstmrs.FullName FROM t_cstmrs CustomerKey INNER JOIN t_prds ON t_prds.ProductKey = t_saldtls.ProductKey INNER JOIN t_saldtls ON t_saldtls.CustomerKey = t_csrmrs. WHERE t_prds.Description = '***Has a lot of the same features as our high end bikes such as a comfortable, durable frame, and high power'***;
Select the items with plush custom saddle and space for panniers	SELECT t_prds.ProductName FROM t_prds WHERE t_prds.Description = ***'The all new plush custom saddle gives you the extra comfort while riding. Newly designed carrier gives you the extra space to add panniers and luggage bags. It provides high stability even when fully loaded.'***
Name the pink products which is aerodynamic bike designed for a woman.	SELECT t_prds.ProductNam e FROM t_prds WHERE t_prds.Description = ***'The sleek aerodynamic body designed for a woman allows you to either race,cross-train, or just socialize. Advanced seat technology provides comfort all day.'*** AND (LOWER(t_prds.Color) = 'pink'
Name the items which are developed by Adventure Works Cycles Professional race team	SELECT t_prds.ProductName FROM t_prds WHERE t_prds.Description = ***'Tt posseses a very light weight heat-treated steel frame, and highly precise steering contro006C. Developed with the Adventure Works Cycles professional race team, is driven only by champions'***

Get the price of items which is a higher end mountain bike with rear suspension	SELECT t_prds.ListPrice FROM t_prds WHERE t_prds.Description = ***'With enhanced performance achieved by using the ground breaking SQ Frame, highly strong rear suspension, this is a higher end model moutain bike’***
Show the price of bikes meant for racing enthusiasts with low budget	SELECT t_prds.ListPrice FROM t_prdcat INNER JOIN t_prdsubcat ON t_prdcat.ProductCategoryKey = t_prdsubcat.ProductCategoryKey INNER JOIN t_prds ON t_prdsubcat.ProductSubCategoryKey = t_prds.ProductSubCategoryKey WHERE t_prds.Description = ***'It is a low budget super strong bike that can ride on all terrains while keeping you in complete control. It is suited for all you racing enthusiasts'*** AND (t_prdcat.ProductCategoryName = 'bikes')
Select an item with mountain wheel for entry-level rider.	SELECT * FROM t_prds WHERE t_prds.Description = ***'Replacement mountain wheel for entry-level rider.'***
Name the items which have pioneering frame technology as the HQ steel frame.	SELECT t_prds.ProductName FROM t_prds WHERE t_prds.Description ***= 'The same pioneering frame technology is used to give you the highest value as the HQ steel frame.'***

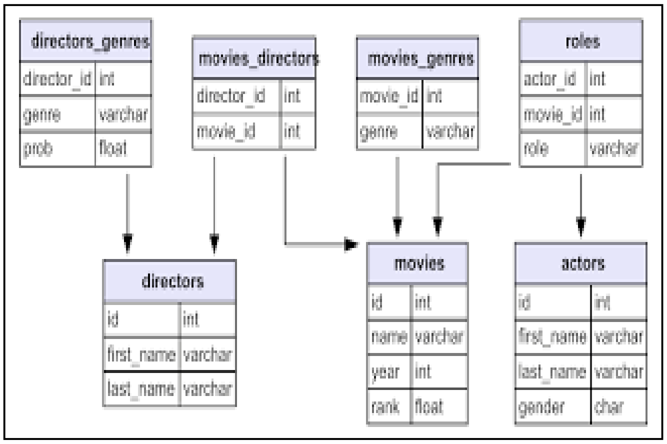


Fig. 7. IMDb Database Schema [21].

TABLE III. IMDB DATABASE: PARTIAL / IMPLIED VALUES

Input Natural Language Query	**Enriched Natural Language Query**	**Output SQL Query**
List the names of actors who played a lawyer	List the names of actors who played a Roles.Role **‘Attorney’**	SELECT actors.first_name FROM actors INNER JOIN roles ON actors.id = roles.actor_id WHERE LOWER(roles.role) = ‘attorney'
List the category of films enacted by Leonardo Dicaprio	List the **movies.genre** of films enacted by actors.first_name ‘Leonardo’ actors.last_name ‘Dicaprio’	SELECT movies.genre FROM movies INNER JOIN roles ON movies.id = roles.movie_id INNER JOIN actors ON actors.id = roles.actor_id WHERE LOWER(actor.first_name) = 'leonardo' AND LOWER(actor.first_name) = 'dicaprio’
Name the satirical movies played by Frank Cady	Name the movies.genre **‘Comedy’** movies played by actors.first_name ‘Frank’ actors.last_name ‘Cady’	SELECT movies.name FROM movies INNER JOIN roles ON movies.id = roles.movie_id INNER JOIN actors ON actors.id = roles.actor_id WHERE LOWER(actor.first_name) = 'frank' AND LOWER(actors.last_name) = ‘cady’ AND LOWER(movies.genre) = ‘comedy'
What are the fairy tale movies directed by Bill Condon	Name the movies.genre **‘Fantasy’** movies played by directors.first_name ‘Bill’ directors.last_name ‘Condon’	SELECT movies.name FROM movies INNER JOIN roles ON movies.id = roles.movie_id INNER JOIN directors ON directors.id = movies_directors.director_id WHERE LOWER(directors.first_name) = ‘bill' AND LOWER(directors.last_name) = ‘condon’ LOWER(movies.genre) = ‘fantasy')
who directed the movie name 10 Rillington Place	who directed the movie movies.name ‘10 Rillington Place’	SELECT directors.first_name , movies.name FROM movies INNER JOIN roles ON movies.id = roles.movie_id INNER JOIN actors ON actors.id = roles.actor_id INNER JOIN movies_directors ON movies.id = movies_directors.movie_id INNER JOIN directors ON directors.id = movies_directors.director_id where movie.name = '10 Rillington Place'
show the movies directed by Richard and enacted by Daniel	show the movies direced by directors.first_name ‘Richard’ and enacted by actors.first_name ‘Daniel’	SELECT movies.name , directors.first_name FROM movies INNER JOIN roles ON movies.id = roles.movie_id INNER JOIN actors ON actors.id = roles.actor_id INNER JOIN movies_directors ON movies.id = movies_directors.movie_id INNER JOIN directors ON directors.id = movies_directors.director_id WHERE directors.first_name = 'Richard' AND actors.first_name = ‘Daniel’

VIII. Conclusion

The proposed system addresses major challenges faced by many of the existing Natural Language to SQL Query conversion algorithms.

Partial and implied data values in the natural language queries are identified by a trained hybrid ML model. WordNet is also used as a safety net to understand implied data values where the vocabulary of the input relational database is not expressive. Descriptive values are identified with the help of Elastic Search.

Using the latent information gathered by the proposed architecture, the accuracy and the robustness of the Natural Language to SQL Query conversion system is proven to increase dramatically (11% to 16%). This has been demonstrated by extensively testing the system on the IMDb database as well as the Company Sales database.

The accuracy of the system is 91.7% on IMDb database and 94.0% on Company Sales database when tested on a diverse set of queries. This is a significantly higher performance compared to the discussed systems in section II.

This system can be used as a plug-in to any of the conversion systems being developed/used. The meta information extracted by the system helps developers boost the accuracy and robustness of their algorithm.

References

[1] Sathick KJ, Jaya A. Natural Language to SQL Generation for Semantic Knowledge Extraction in Social Web Sources. Indian Journal of Science and Technology. 2015; 8(1):1–10.

[2] F. Li and H. V. Jagadish. Constructing an interactive natural language interface for relational databases. PVLDB 2014, 8:73–84.

[3] IMDb data files available for download [Online]. Available: https://datasets.imdbws.com [Accessed: 19- Mar- 2020].

[4] Navid Yaghmazadeh, Yuepeng Wang, Isil Dillig, and Thomas Dillig. 2017. SQLizer: query synthesis from natural language. Proc. ACM Program. Lang. 1, OOPSLA, Article 63 (October 2017), 26 pages. DOI:https://doi.org/10.1145/3133887.

[5] N. Ott, "Aspects of the automatic generation of SQL statements in a natural language query interface", Information Systems, vol. 17, no. 2, pp. 147-159, 1992.

[6] Xiaojun Xu, Chang Liu, and Dawn Song. 2017. Sqlnet: Generating structured queries from natural language without reinforcement learning. arXiv preprint ArXiv:1711.04436.

[7] Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2SQL: Generating structured queries from natural language using reinforcement learning. arXiv Preprint arXiv:1709.00103 (2017).

[8] Giordani, A. and Moschitti, A. (2009b). Syntactic structural kernels for natural language interfaces to databases. In Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part I, ECML PKDD, Berlin, Heidelberg. Springer-Verlag.'09, pages 391–406.

[9] Prasetya Utama, Nathaniel Weir, Fuat Basik, Carsten Binnig, Ugur Cetintemel, Benjamin Hättasch, Amir Ilkhechi, Shekar Ramaswamy and Arif Usta.An End-to-end Neural Natural Language Interface for Databases. arXiv Preprint arXiv:1804:00401.

[10] A. Prasad, G. Shobha, N. Deepamala, S. S. Badhya, Y. Yashwanth and S. Rohan, "Machine Learning Techniques to Understand Partial and Implied Data Values for Conversion of Natural Language to SQL Queries on HPCC Systems," 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), Bengaluru, India, 2019, pp. 1-5. doi: 10.1109/CSITSS47250.2019.9031035.

[11] S. S. Badhya, A. Prasad, S. Rohan, Y. S. Yashwanth, N. Deepamala and G. Shobha, "Natural Language to Structured Query Language using Elasticsearch for descriptive columns," 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), Bengaluru, India, 2019, pp. 1-5. doi: 10.1109/CSITSS47250.2019.9031030.

[12] Apache Commons CSV API Documentation from the official website. https://commons.apache.org/proper/commons-csv/apidocs/index.html [Accessed: 19- Mar- 2020].

[13] R. Ruizendaal, "Deep Learning #4: Why You Need to Start Using Embedding Layers", Medium, 2019. [Online]. Available: https://towards datascience.com/deep-learning-4-embedding-layersf9a 02 d55ac12. [Accessed: 19- Mar- 2020].

[14] S. Yan, LSTM node. 2016. [Online]. Available: https://medium.com/ mlreview/understanding-lstm-and-its-diagrams-37e2f46f1714. [Accessed: 19 - Mar- 2020].

[15] Bulk API, Elasticsearch Reference (7.2) in the official website. https://www.elastic.co/guide/en/elasticsearch/reference/current/docsbulk .html [Accessed: 19- Mar- 2020].

[16] Bulk Helpers, Python Elasticsearch Client in the official website. https://elasticsearch-py.readthedocs.io/en/master/helpers.html [Accessed: 19- Mar- 2020].

[17] Stop Analyzer, Elasticsearch Reference (7.2) in the official website. https://www.elastic.co/guide/en/elasticsearch/reference/current/analysi s-stop-analyzer.html [Accessed: 19- Mar- 2020].

[18] Language Analyzers, Elasticsearch Reference (7.2) in the official website. https://www.elastic.co/guide/en/elasticsearch/reference/current/ analysi s-lang-analyzer.html [Accessed: 19- Mar- 2020].

[19] Multi-match query, Elasticsearch Reference (7.2) in the official website. https://www.elastic.co/guide/en/elasticsearch/reference/current/querydsl-multi-match-query.html [Accessed: 19- Mar- 2020].

[20] TFIDFSimilarity (Lucene 4.6.0 API) in the official website. https://lucene.apache.org/core/4_6_0/core/org/apache/lucene/search/ similarities/TFIDFSimilarity.html [Accessed: 19- Mar- 2020].

[21] IMDb Dataset Schema [Online]. Available: http://kt.ijs.si/ janez_kranjc/ilp_datasets [Accessed: 19- Mar- 2020].

Digital Transformation in Higher Education: A Framework for Maturity Assessment

Adam Marks[1], Maytha AL-Ali[2], Reem Atassi[3], Abedallah Zaid Abualkishik[4], Yacine Rezgui[5]
American University of the Emirates[1, 3, 4]
Zayed University, Dubai, UAE[2]
University of Cardiff, Cardiff, UK[5]

***Abstract*—Literature in digital transformation maturity is scarce. Digital transformation in higher education, especially after COVID-19 is seen as inevitable. This research explores digital transformation maturity and challenges within Higher Education. The significance of this study stems from the role digital transformation plays in today's knowledge economy. This study proposes a new framework based on Deloitte's 2019 digital transformation assessment framework with Petkovic 2014 mega and major higher education process mapping. The study triangulates the findings of multiple research instruments, including survey, interviews, case study, and direct observation. The research findings show a significant variance between the respondents' perception of digital transformations maturity levels, and the core requirements of digital transformation maturity. The findings also show the lack of holistic vision, digital transformation competency, and data structure and processing as the leading challenges of digital transformation.**

Keywords—Digital transformation; higher education; COVID-19

I. Introduction

Organizations today are operating in a massive, digitally connected world, and their stakeholders expect seamless and personalized digital services [2], and [9]. Producing and acquiring knowledge has a great importance today. The success of organizations and nations highly depends on producing and using information successfully. The increased use and production of knowledge places organizations into a necessary digital transformation. This digital transformation impacts the core components of an organization - from its operating model to its infrastructure. Organizations usually do not transform by choice, more often when they fail to evolve and keep up with market changes and technology disruptions [24].

Terms like digitization, digitalization, and digital transformation can be confusing, especially if used interchangeably, however they refer to distinct concepts. While digitization is concerned with transforming analog objects into digital representations, digitalization is concerned with improving processes by use of digitized data and programs, also known as automation [10]. Digital transformation is concerned with transforming organizational processes; build new competencies and models through digital technologies in a profound and strategic way [12]. Digital transformation refers to an organizational change realized by means of digital technologies and business models with the aim to improve organization's operational performance. It involves much more than implementing a well-chosen technology solution, it is a close alignment between information technology and business processes that will lead to a substantial outcome for the organization, keeping in mind organizational readiness, change management, and managing key stakeholders [15].

The author in [4] states "Digital transformation is not only the transformation of tool, technologies and process but it is transformation of entire business model. It changes the way a business operates and interacts within itself and with external world. Business transformation is a change in mind set that business is evolving faster than we are adapting". Digital transformation should have a strategic objective defined in an operational architecture with business use cases.

Similar to other industries, higher education institutions need to make informed decisions, and at times quick decisions to streamline operations, understand their customers, service delivery, product development, asset utilization or other operational areas. Data should be used to discover hidden patterns, and underlying performance in specific areas, and influence decisions that yield maximum impact for the organization. Legacy systems can result in significant cost and strain budgets. Moreover, threats to higher institutions, both online and on campus, present an urgent need for security and intelligence about students and staff more than ever before. Those challenges cannot be addressed using manual forms and processes. Higher education institutions today must integrate digital technologies into their business to a much greater extent than before [12] and [15].

The impact and magnitude of the COVID-19 pandemic forced many sectors to attempt to do business online. Education and higher education institutions across the globe had to make quick and important digital transformation adjustments to sustain operations. Questions about course delivery, virtual classrooms, seats, capacity, conducting exams and assessments, academic integrity, use of web cams, capacity and quality of video conferencing, and many other questions were raised. Many universities signed up with Zoom, MS Teams, Respondus, and other software systems to sustain operations.

Literature in digital transformation maturity and challenges, specifically within higher education, and more specifically within developing nations is scarce. This study aims to address those identified gaps. Given the importance of higher education in today's information society and knowledge economy, this study is significant to higher education institutions, as well as to other stakeholders involved in the hierarchys of higher education, including students, educators, researchers, institutions, and government agencies. Digital transformation in higher education, especially after COVID-19 is seen as inevitable not only to compete, but also to survive and sustain key operations.

This research explores digital transformation maturity and challenges within United Arab Emirates "UAE", one of the advanced developing nations in terms of IT infrastructure and digital transformation plans. The significance of the study does not only stem from the critical role higher education is responsible in educating and training future leaders, workers, and citizens; but also from the key role digital transformation plays in today's knowledge economy, which became more evident after the COVID-19 pandemic.

The next section of this paper presents some of the literature related to the main topic of this study, followed by the research methodology, the discussion, conclusion, and recommendation.

II. Related Work

Digital transformation is a key element of the 4th industrial revolution. The author in [20] describes the emergence of the fourth industrial revolution by linking three fundamental factors. These are: "Speed: New technologies that are connected to each other and are very versatile move quickly at an exponential speed, triggering each other. Width and Depth: Digitization speeds up the industry 4.0. However, the increase in technology diversity in the industry has brought about the change. System Impact: Industry 4.0 is expected to undergo a total change as digital industries, companies, and even countries".

The author in [13] breaks digital transformation into three attributes: creating value, optimizing the processes that execute a vision of customer experiences, and building foundational capabilities that support the entire structure. Fig. 1 shows the three attributes of digital transformation with more details.

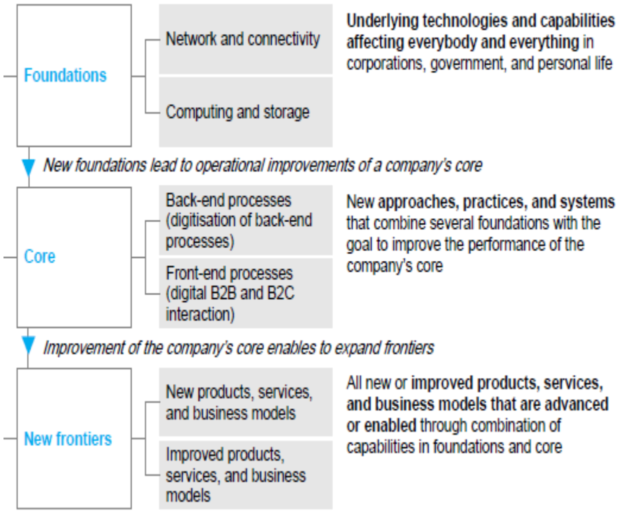

Foundations
Network and connectivity
Computing and storage
Underlying technologies and capabilities affecting everybody and everything in corporations, government, and personal life
New foundations lead to operational improvements of a company's core
Core
Back-end processes (digitisation of back-end processes)
Front-end processes (digital B2B and B2C interaction)
New approaches, practices, and systems that combine several foundations with the goal to improve the performance of the company's core
Improvement of the company's core enables to expand frontiers
New frontiers
New products, services, and business models
Improved products, services, and business models
All new or improved products, services, and business models that are advanced or enabled through combination of capabilities in foundations and core

Fig. 1. Digital Transformation Attributes - [13].

There are several reasons why organizations undergo digital transformation; however, the main reasons are related to the issues of competitive advantage and survival. Digital transformation of an organization represents an objective process capable at responding to disruption in critical functions and changing organizations environments [21], and [23]. As in other industries, four elements are driving digital transformation in education: customer experience, competitiveness, profitability, and agility [11].

The impact of digital transformation transcends technology. Through the process of digital transformation, organizations use multiple new digital technologies, with the intent to achieve superior performance and sustained competitive advantage. In such way, they transform different dimensions of business, such as the business model, the customer experience and operations, and simultaneously impacting people and networks [13].

The author in [6] presents six dimensions to digital transformation, they are: Established and accepted organizational digital strategy, Organized agile, and adaptable collaborative processes in modern business models; Complete automation of business processes; Detailed analysis and research of customers' decision making; Information technology supporting all organizational business processes; Usable and relevant data, use of data analytics as a basis for decision making in line with the organization's goals and strategy.

The author in [3] talks about four dimensions of digital transformation, namely "the purpose", "degree of strategy", "speed of strategy", and "the value source". Table I shows those dimensions in more details.

Fig. 2 outlines the main sectors that have been disrupted by digital transformation. Public Sector related entities, including education, are being disrupted now by digital transformation. The opportunity for the Public Sector is to learn from the previous experiences of other sectors.

TABLE I. Digital Transformation: Dimensions, Issues, and Implications [3]

Dimension of Digital Transformation	Questions for Manager (Strategy, organization, and Business Models)	Main Topics
The purpose of digitation strategy	Which analytical methods will be selected in the company What are the spaces for development and value creation	Determining and analyzing the value creation space
Degree of digitization strategy	What is the relative importance of platforms? What kind of typology? Which governance structure promotes innovation?	Defining and analyzing the idea of creating new platforms
The speed of digitization strategy	How to define innovation offers	Fast and systematic phenomena
Value sources, creation based on digital strategy	What are the sources of value creation in digital space	Define the proposed values of the digital space

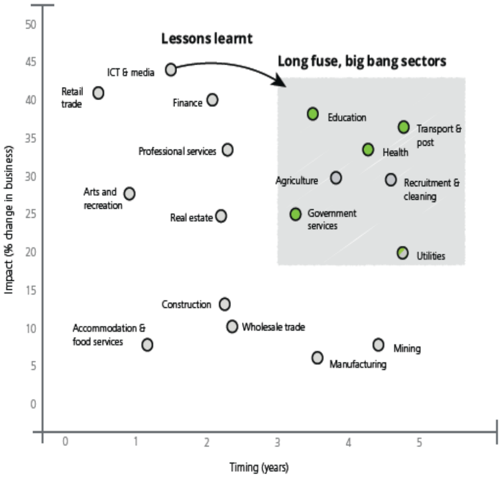


Fig. 2. Sectors affected by Digital Transformation – [5].

In the Middle East and North Africa "MENA" region, the trends of information and communications technologies (ICT) are very diverse due to different levels of development both between and within countries. This gap can be attributed to different aspects including infrastructure, economic conditions, job market, and lack of adequate governance. Nevertheless, nearly all countries in the region are pursuing policies supporting digitalization to further development.

Countries – such as the United Arab Emirates (UAE), and Saudi Arabia, are well equipped for further technological development [8]. However, the National ICT Index still shows that those countries still lag behind other developed economies in terms of Digital Government capabilities [13] and [17].

To address their own economic, social and environmental challenges, several governments in the MENA region have launched national transformation plans with a focus on enabling ICT and Digital Transformation technologies. Integrating digital technologies to develop smarter cities and become smarter nations is a key outcome of their national visions [13] and [15].

Globally, higher education digital transformation is highly influenced by government policies and institutional development strategies [26]. The aim of the digital transformation process in higher education is to redefine educational services and redevelop operational processes. This can be achieved using three possible approaches. The first approach involves service-first transformation. It focuses on changing and redefining services before making key improvements and changes to operations. The second approach is the operation-first transformation. In this approach, the higher education institution identifies new and amends current digital processes, activities and operations. The third approach is service-operation combination, involves integrated transformation through systematic interrelation of both previous approaches [19].

As shown in Table II, [18] maps higher education business processes into four hierarchical levels: mega processes, major processes, sub-processes, activities and tasks. The mega processes include the learning and teaching process, the research process, the enabling process, and the planning and governance process.

The author in [5] presents a Digital Maturity Assessment Framework, using five key criteria:

- Does the organization have the right vision and strategy for digital, and the leadership, communications and focus required to support this vision?
- Does the organization have the right talent, skills and knowledge to support its vision, products, and services?
- Does the organization have the right processes, controls and digital technologies to support the operations of the organization?
- Does the organization have the right technologies and infrastructure as well as the ability to develop, manage and deliver?
- Does the organization have the right approach to understanding and communicating with its customers to succeed in a digital environment? [12]

The author in [16] study provide a similar presentation of higher education business processes and example of new digital trends, as shown in Table III.

Fig. 3 visualizes results from Gartner's 2017 CIO Survey, shows where higher education intuitions are in terms of digital transformation. Organizational mind-set is what separates the No digital initiative institutions from those with Desire/Ambition. This transformation requires a framework that is effectively communicated to key stakeholders and decision makers. If employed correctly, digital transformation can play a major role in today's higher education including in the areas of admission tracking, enrollment optimization, and academic advising [7].

TABLE II. OVERVIEW OF MEGA AND MAJOR PROCESSES IN HIGHER EDUCATION INSTITUTIONS [18]

Learning and Teaching Process	**Research Process**
Study program accreditation	Research planning
Teaching process preparation and realization	Research preparation
Teaching process outcomes monitoring	Research conduct
Teaching process assessment	Research outcomes monitoring
Student and teacher mobility realization	Research evaluation
Enabling Process	**Planning and Governance Process**
Student administration services	Organization management services
Library services	Change and business process management
Staff provision and development services	Plan development
Finance and accounting services	Budget ad funds planning
Marketing, sales, and distribution services	Performance assessment
Procurement services	

TABLE III. CATEGORIES OF EDUCATIONAL SERVICES AND NEW DIGITAL TRENDS [16]

Categories of educational services	**Examples of new digital trends**
Administration	Application for enrollments, enrollments for exams, grade generation, class schedule
Communication	Collaboration platforms, online communication
Teaching and preparing lessons	Electronic books, online learning resources
Teaching and learning	Online learning
Reviews and examinations	Reviewing test and exams, Exam grading

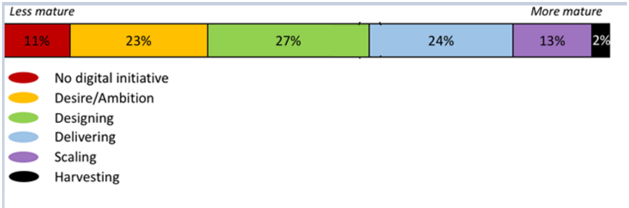


Fig. 3. Digital Transformation Maturity in Higher Education- Source: [7].

The literature shows that digital transformation is usually faced with a number of challenges. More often than not, those challenges are not listed in any specific order based on criticality, and they are not attached to a specific industry. Challenges reported include: the changing customer experience/expectations, resistance to change, resistance to technology, lack of leadership support, lack of competency and digital transformation skills, failing or poor analytics, lagging and legacy business models and systems, poor planning, misalignment with business strategy, technology and data challenges, lack of clear vision, and digital literacy of stakeholders [18], [14], and [11].

If implemented correctly, digital transformation tools and technologies such as Artificial Intelligence (AI), Internet of Things (IoT), big data, block chain, social analytics, and cloud services can enhance and change educational system practices, especially in a world where students are constantly interacting with technology in all other aspects of their everyday lives, digital transformation clearly offers opportunities for student engagement that are not always available in the fixed environment of the classroom [1], and [27].

The author in [22] discuss four ways where digital transformation may change how education looks in the future, namely smart content, differentiated and personalized learning, global and remote learning, and administrative efficiencies.

In the area of "Smart Content", digital transformation can be attained through e-books, new learning platforms, virtual content such as video lectures and conferences, electronic curricula, distributed educative information across devices. Similarly, in the area of "Differentiated and Personalized Learning", digital transformation can be attained through personalized electronic tutoring customized to the learning styles and particular needs of the student. The traditional curriculum is designed to suit as many students as possible. For students in the top 10 percent and the bottom 10 percent, AI for example can be used to provide testing and feedback to those students to give them challenges they are ready for, identify gaps in knowledge and re-direct them to new topics when appropriate.

In the area of "Global and Remote Learning", digital transformation can facilitate learning from anywhere and at any time. Furthermore, it can be used to support students with homework and exam preparation remotely with advanced tutoring and study programs.

Last, in the area of "Administrative Efficiencies", digital transformation can support higher education to shift from wondering about the future into predicting, forecasting, and shaping the future; making proactive and informed decisions and taking action based on that information. Digital transformation can support universities in using conventional and unconventional (unstructured), internal and external data to discover hidden patterns underlying performance in different areas, track admissions, optimize enrolment, manage grants, enhance academic advising etc. Digital transformation can help higher education institutions to know what is happening (descriptive analytics), what is likely to happen in the future (predictive analytics) and to examine trends, causes and likely forecasts and use that information to make decisions (prescriptive analytics) [22], and [25].

III. METHODOLOGY

While the importance of digital transformation is recognized, data about digital transformation maturity and challenges is scarce, especially in developing nations, and specifically within higher education. This study explores digital transformation maturity level in Higher Education Institutions in UAE higher education. The study uses a new framework that is based on the Petkovic's 2014 mega and major processes mapping, and [12] maturity assessment framework. The researchers believe that Petkovic provides a balanced and comprehensive classification of higher education business processes using four hierarchical levels: mega processes, major processes, sub-processes, activities and tasks. Unlike other classifications, the Petkovic's classification does not contain overlapping, ambiguous, and repetitive processes. The researchers also believe that digital transformation maturity criteria listed in the maturity assessment framework by [12] are comprehensive, tangible, easy to understand and reflective. Based on the above, the combination of both models provides a good starting point for higher education institution to assess their digital transformation maturity level, and identify areas that need improvements. The proposed framework in this study is flexible, customizable, and can support further more detailed analysis as required. The study examines public and private higher education institutions in the UAE. It ranks the criticality of digital transformation challenges using identified pattern codes such as the regulatory and business environment, IT infrastructure, data governance, affordability and budget constraints, personnel competency and IT skills, etc. We consider that the problem identification and related key issues are very important in order to achieve successful implementation of digital transformation.

The goal of this study is achieved throughout several objectives, beginning from the literature study of the state of the art, continuing with the wide-distributed survey, in-depth semi-structured interviews, direct observation of the researchers, and case study. The literature study provided the possibility of identifying and analyzing trends related to the topic, while the survey, interviews, direct observation, and case study provided the possibility of identification and analysis of trends in the field of digital transformation at the national level in public and private higher education institutions. This study is a phenomenological research to determine the views of IT director and senior academicians on the maturity and challenges of digital transformation. Phenomenological researches may not reveal generalizable situations, but they can provide examples, explanations and experiences that will help to a phenomenon identified and understood better [25], and [15].

The survey was conducted in both public and private higher education institutions, targeting IT directors, chief information officers, and senior academicians concerned with digital transformation. The survey was sent to 61 individuals. Response was received from 52. The questionnaire design and construction consist of 15 Likert Scale closed-ended multiple-choice-five-pointer questions. Respondents were required to complete the questionnaire indicating the extent to which they agree or disagree with the questions. A room for comments for each question was also available. The survey questions were divided into three sections; the first section measures the respondent's view on the institution's level of digital transformation maturity; the second section verifies the existence/non-existence of key elements of digital transformation maturity; and the third section ask about the respondents rating of the challenges faced.

Six in-depth, semi-structured interviews were held with IT directors, and another four interviews were held with senior academic administrators to gain deeper understanding of expected value and the challenges faced during digital transformation; direct observation was used to verify what people do, rather than what people say they do, lastly, a case study was conducted at one of the public universities to validate and triangulate the results of the survey, direct observation, and the interviews.

The importance and the rationale of this research lie in the identification and analyses of the readiness of higher education institutions in the UAE to embrace a meaningful and mature level of digital transformation in higher education academic processes post COVID-19.

The research results can be used as an important input to the design of new academic processes that would be more effective, aligned, efficient, and cost-effective. Moreover, this study points to the key challenges faced by higher education institutions in the UAE in achieving mature digital transformation, turning data into a valuable asset that could be used for prescriptive, corrective and predictive decision-making, using a proposed framework to measure digital transformation maturity, and pinpoint areas of concern and areas of strength. The research also provides the practitioners from the field of digital technologies with the information and knowledge related to their potential market and the related trends.

IV. Discussion

In this section, we present the findings of this study, responding directly to the two key research questions.

What is the level of digital transformation maturity in the examined UAE higher education institutions?

The data collected in this study shows a significant variance between digital transformations maturity level perception reported by the respondents, and the core requirements of digital transformation maturity. While more than 80 percent of the examined institutions reported a digital transformation maturity level between "delivering or harvesting" as seen in Fig. 4, none of them had a comprehensive digital transformation plan.

A few reported a list of digital transformation initiatives. However, those list of initiatives were more aligned with automation not digital transformation, and they were mostly initiated to comply with external regulatory requirements by the Telecommunication Regulatory Authority TRA. In almost all cases, digital transformation initiatives had no connection to a return on investment, gained value, or a real transformation of a business process. The great majority of cases were concerned with the automation of electronic forms, adding workflow and approvals. None of the initiatives reported was concerned with analytics, machine learning, artificial intelligence, big data and other more-advanced digital transformation technologies.

As shown in Table IV, based on proposed higher education digital transformation maturity assessment model, it was noticed that most institution focused their digital transformation effort in the area of enabling processes, while much less in planning and governance, learning and teaching, and research respectively (see Table II). Respondents attributed this to the availability of third party systems supporting student administration, library services, finance an accounting, etc. Ellucian and Oracle are among the main contenders in this area. Systems supporting accreditation, research, and curriculum management are relatively new in the market, compared to systems supporting enabling processes. Finally, the data findings shows clearly how critical is the role of IT governance in ensuring that all mega and major processes receive the needed support. A segmented structure of data ownership can deliver a segmented vision that can directly affect digital transformation completeness, effectiveness, and alignment.

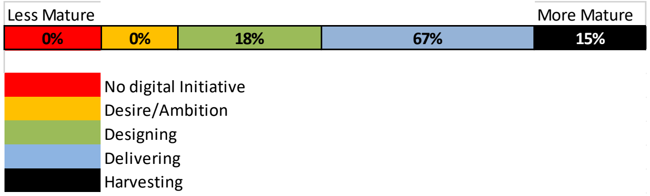


Fig. 4. Digital Transformation Maturity reported.

TABLE IV. Reported DT Maturity Perception by Mega and Major Process

Higher Education Mega and Major Processes	**DT Maturity**
Learning and Teaching Processes	
Study Program Accreditation)	32%
Teaching processes preparation and realization	74%
Teaching process Assessment	22%
Student and Teacher Mobility Realization	38%
Enabling Processes	
Student Administration Services	92%
Library Services	86%
Staff provision and development services	65%
Finance and accounting services	83%
Marketing, sales, and distribution services	56%
Procurement services	83%
Research Processes	
Research planning	36%
Research preparation	32%
research conduct	18%
Research outcome monitoring	18%
Research evaluation	24%
Planning and governance processes	
Organization management services	54%
Change and business process management	42%
Plan development	64%
Budget and fund planning	88%
Performance assets	36%
Higher Education Mega and Major Processes	

TABLE V. Reported DT Challenges in UAE Higher Education

Digital Transformation Challenge	**Consensus**
Wholistic Vision	76%
Personnel Competency and IT Skills	54%
Data Structure, Data Processing, and Data Reporting	52%
Redundant Systems	42%
Third Party Reporting Systems	42%
Manual Entries (Middle Man)	38%
Potential Use by Customers	28%
Regulatory and Business Environment	16%
Social and Economic Impact	12%
Privacy and Security Concerns	4%
IT infrastructure	3%
Affordability and Budget Constraints	2%
Other capability constraints	1%

What are the key digital transformation challenges facing the UAE examined higher education institutions?

The data findings of this study reveals a number of digital transformation challenges. While some of the reported challenges are consistent with some of the challenges reported in literature, some of the challenges listed were specific to the UAE higher education. Table V lists digital transformation challenges in UAE higher education, as reported by the respondents.

A. *Holistic Vision*

The data supported findings of this study show that the most critical challenge facing digital transformation in UAE higher education is the lack of a holistic digital transformation vision. The data reveal that none of the examined institutions had a stand-alone digital transformation vision or plan. Two public universities had list of key performance indicators to satisfy TRA requirements. Most of the other institutions worked on a list of ad-hoc automation initiatives, mainly driven by IT personnel. In several cases, redundant processes and shadow systems were identified. Some of the respondents reported confusion about digital transformation ownership. Several IT directors, chief information officers, and academic administrators did not believe that they are responsible for digital transformation. IT directors and chief information officers expected academic administrators to identify, initiate, and prioritize what processes should be digitally transformed, while academic administrators viewed digital transformation as a technical process that should be driven, initiated, and prioritized by IT directors and the chief information officers. Some of the respondents viewed digital transformation as a joint responsibility that lacks proper monitoring, initiation, and management.

B. *Personnel Competency and IT Skills*

The second critical challenge of digital transformation reported was the lack of critical IT skills. IT Personnel in the UAE are mainly comprised of Asian expatriates. IT personnel lack previous experience in the higher education sector, and support English-based systems, English-based curricula, and operations. Similarly, many IT managers and directors did not have prior higher education or technical experience, which is critical in establishing a holistic digital transformation vision and plan.

Respondents reported systems that were not used, while annual license is regularly paid; in-house developed systems that were redundant; IT governance, that was not established, requests for new systems driven by individuals based on familiarity; segmentally initiated digital transformation decisions; poor and underdeveloped automation efforts that did not create any real value for the institution or provide services with customer –centric view.

Given the profile of the workforce in the UAE, this digital transformation challenge is more specific to the UAE higher education environment. Experience with systems such as Banner, people soft, campus solution, etc. is more difficult to attract in the MENA region compared to the US and Europe. With few exceptions, many universities are supporting critical operations, including admission, registration, advising, scheduling, and graduation with personnel that have had no prior experience with those systems or higher education.

C. Data Structure, Data Processing, and Data Reporting

The Third critical challenge reported was data structure, processing, and reporting (input-process-output). This challenge can be linked to the second reported challenge, as it is also more pertinent to the UAE higher education environment.

Because of the lack of fundamentals such as an IT strategy, IT governance, and/or data governance, data structures and key codes were not setup correctly, and key modules were not utilized. For instance, one institution used Banner Student Information Systems "SIS", but did not utilize Banner workload module. Instead built a stand-alone system to manage faculty workload. No one knew that a small number of forms need to be populated and a fully integrated module will be available in a more efficient way.

Inconsistent college codes, program codes, major codes, etc. used across different creating inconsistent data outputs that is very difficult to verify. Business rule codes were also not well defined, organizations ended up with problems in critical academic, business, and financial areas, including major out of balance issues. Data reporting relied heavily on data extraction and ad-hoc (backend) queries and reporting. System built-in reports were limited or short of customer's specific needs; keeping in mind that most systems are made to align with the US higher education environment including Ellucian, Blackboard, Leepfrog, CurriCUNET, Taskstream, Oracle, etc. This variance in input-processing-output created a major hurdle for integration, consistency, and reporting, resulting in failure to create value and enable an effective digital transformation strategy.

D. Redundant Systems

The fourth challenge facing higher education institutions in UAE was the existence of redundant systems. As referred to earlier, due to the lack of proper IT and data governance, several organizations did not have a proper system/software acquisition process in place, subsequently resulting in systems, functions, and data overlap and redundancy, creating major challenges for integration and data consistency, reliability, integrity, availability, timeliness, and confidentiality.

E. Third-Part Reporting Systems

The fifth challenges cited by higher education institutions attempting to digitally transform their processes was the existence of several external reporting agencies/systems that require different data sets, formats, and requirements, including different accreditation reporting systems, and other compliance reporting systems.

Several institutions felt that need to manually extract the required data from different internal systems, then format the data sets as required for reporting purposes. Microsoft Excel formatted files are widely used to support this function.

F. Manual Entries (Middle Man)

Because institutions were required to report to multiple external entities using third-party reporting systems, a lot of data extraction, data manipulation, data formatting, and data entry was taking place; in several cases the people responsible for data extractions from the organization's different systems, were completely different from those making manual entries into third party systems, potentially leading to system and data and submission errors.

G. Potential use by Customer (Adoption)

The seventh challenge facing higher education institutions in UAE was the potential use by customers. Several respondents cited resistant to change, resistant to technology, buy-in, awareness, and training as leading causes for adoption challenges. Some of the respondents stated that processes were detached from systems causing loopholes, delays, redundancy, and errors. Other respondents cited off-the shelf systems as generic, while others did not support in-house developed system by IT personnel, and viewed them as temp-cheap solutions, driven by the insecurity of IT personnel to guarantee their jobs.

H. Regulatory and Business Environment

The eighth challenge extracted from the data set was the regulatory and the business environment. Private institutions felt more at liberty than government institutions in this category. This is natural, given that government institutions receive full funding from the government, and the government audits their books. In addition, the Telecommunication Regulatory Agency TRA has its own protocols and requirements that must be observed, including what data can or cannot be on the cloud. The purchasing process in government institutions also has its own special requirements that may linger or hinder the process of acquiring certain IT assets that would support digital transformation.

I. Social and Economic Impact

Although was only cited by 12% of the respondents, the ninth challenge of digital transformation was concerned with the social and economic impact. Some higher education institutions reported that some of the digital transformation initiatives were not rolled out due to social concerns about how the institution will be viewed, including cases where different genders may have direct communications or use of web cams. Many institutions stressed the importance of showing respect of the UAE culture and norms as one of the main factors for attracting UAE national students.

J. Privacy and Security Concerns

Only cited by 4% of the respondents, privacy and security concerns was cited as the tenth challenge to digital transformation. Some universities did not feel that their hardware, security, and network was not ready to manage potential security threats that could come with the expansion of their digital infrastructure.

K. IT Infrastructure

While the majority of respondents viewed their universities IT infrastructure as mature and ready to support digital transformation, 3% of the respondents expressed concerns about the full readiness of their IT infrastructure in its entirety, and reported it as the eleventh challenge.

L. Affordability and Budget Constraints

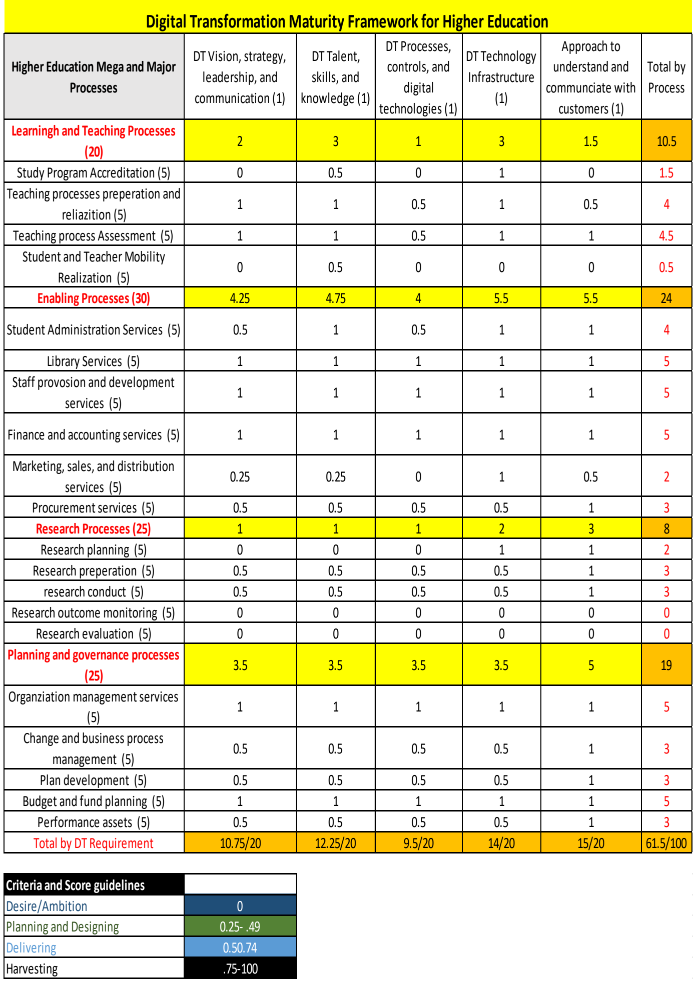
Surprisingly only 2% of the respondents reported affordability and budget constraints as the twelfth challenge facing digital transformation in their organizations. Understandably, those were from small private universities, not government universities.

M. Other Capability Consraints

The final and the thirteenth challenge reported was reported by only 1%, and it was concerned with random capability reasons that were not directly identified, but simply reported as capability constraints.

The Proposed Higher Education Digital Transformation Maturity Assessment Framework: The data findings of this study shows a significant variance between the respondents' digital transformation maturity perception, and the core criteria of digital transformation maturity. Moreover, higher education is faced with a number of digital transformation challenges. A higher education digital transformation maturity assessment framework can provide higher education institutions with guidance, criteria, and an assessment of strength and weakness areas, mapped to mega and major process.

The framework proposed in this study combines the mapping of higher education processes with the digital maturity assessment framework to create an assessment framework to measure digital transformation maturity in higher education. For the sake of illustration, Fig. 5 shows an example where the framework is used with equal weight assigned to each mega process and each major processes. Institution may choose to assign different weights. The framework acts similar to a scorecard, pinpointing areas of strength and areas of weakness across each mega and major processes, and across each maturity criterion. Using the example below, it is easy to see that the research process and the teaching and learning process are the weakest in digital transformation. Users can drill down further to see that research monitoring and evaluation are among the lowest scores. Similarly, one can also see looking at the maturity requirements that vision, strategy, and processes, and controls are among the lowest criteria. The framework can help higher education institutions track their digital transformation progress and benchmark it regularly. Institution can choose to go from mega and major processes into tasks and activities if they wish to add more details.



Digital Transformation Maturity Framework for Higher Education

Higher Education Mega and Major Processes	DT Vision, strategy, leadership, and communication (1)	DT Talent, skills, and knowledge (1)	DT Processes, controls, and digital technologies (1)	DT Technology Infrastructure (1)	Approach to understand and communciate with customers (1)	Total by Process
Learningh and Teaching Processes (20)	2	3	1	3	1.5	10.5
Study Program Accreditation (5)	0	0.5	0	1	0	1.5
Teaching processes preperation and reliazition (5)	1	1	0.5	1	0.5	4
Teaching process Assessment (5)	1	1	0.5	1	1	4.5
Student and Teacher Mobility Realization (5)	0	0.5	0	0	0	0.5
Enabling Processes (30)	4.25	4.75	4	5.5	5.5	24
Student Administration Services (5)	0.5	1	0.5	1	1	4
Library Services (5)	1	1	1	1	1	5
Staff provosion and development services (5)	1	1	1	1	1	5
Finance and accounting services (5)	1	1	1	1	1	5
Marketing, sales, and distribution services (5)	0.25	0.25	0	1	0.5	2
Procurement services (5)	0.5	0.5	0.5	0.5	1	3
Research Processes (25)	1	1	1	2	3	8
Research planning (5)	0	0	0	1	1	2
Research preperation (5)	0.5	0.5	0.5	0.5	1	3
research conduct (5)	0.5	0.5	0.5	0.5	1	3
Research outcome monitoring (5)	0	0	0	0	0	0
Research evaluation (5)	0	0	0	0	0	0
Planning and governance processes (25)	3.5	3.5	3.5	3.5	5	19
Organziation management services (5)	1	1	1	1	1	5
Change and business process management (5)	0.5	0.5	0.5	0.5	1	3
Plan development (5)	0.5	0.5	0.5	0.5	1	3
Budget and fund planning (5)	1	1	1	1	1	5
Performance assets (5)	0.5	0.5	0.5	0.5	1	3
Total by DT Requirement	10.75/20	12.25/20	9.5/20	14/20	15/20	61.5/100

Criteria and Score guidelines	
Desire/Ambition	0
Planning and Designing	0.25- .49
Delivering	0.50.74
Harvesting	.75-100

Fig. 5. Higher Education Digital Transformation Maturity Framework.

V. Conclusion

Digital transformation is one of the biggest catalysts of the business environment today, and higher education is not excluded from this evolution. It is a move that goes beyond the scope of systems and new technologies, while also representing the modernization of organization philosophy, purpose, competition, and patterns that change with emerging audiences. As the business environment, students, and employees change, they do so at an accelerated speed that often exceed an organization's ability to adapt. This disruption causes critical business functions and processes to inevitably be exposed within and outside the organization; subsequently requiring the restoring of new investments in technology, business models and processes to more effectively compete in a continual digital economy shift. Digital transformation is an inevitable choice for higher education institutions everywhere, especially after COVID-19.

Digital transformation is a process that can hardly be historically compared to any other process, as it does not exclude the development levels of different countries.

In other words, all countries, regardless of their development level must undergo some level of digital transformation; and while in the developed world, the need for digital transformation has been reinforced and installed, and organizations and governments have developed sophisticated methods for applying digital technology to create products or to deliver certain services, and add value, some developing countries are still attempting to move from desire and ambition to planning, delivering and harvesting.

Despite all the talk about digital transformation in developed and developing countries, and across all industries, the reality is that digital transformation is only as useful as its rate of true implementation and return on investment. Otherwise, organizations will not benefit in terms of efficiency, effectiveness, cost-savings, competitive advantage, and decision-making.

For a number of decades, higher education institutions globally claimed digital transformation maturity, citing students' information systems, learning management systems, etc. The COVID-19 pandemic forced many institutions to use remote teaching, disrupting the regular and normal business environment and operations, subsequently exposing critical functions and their true level of digital transformation maturity and challenges.

The UAE is one of the leading developing nations in terms of IT infrastructure, and the adoption of new technologies. The UAE government has made significant leaps in e-government, e-commerce, e-business, and e-services in general. There are several agencies contributing to this advancement, including the Ministry of Artificial Intelligence, Smart Dubai, and the Telecommunication Regulatory Authority.

Given the importance of digital transformation, higher education, and the role they both play in today's digital/knowledge economy, the aim of this study is to examine the digital transformation maturity level and challenges in UAE higher education institutions in the after math of COVID-19, and the need to provide remote e-service to students, employees, and other customers and stakeholders.

The first research question is concerned with measuring the level of digital transformation maturity in UAE higher education institution using Deloitte's digital maturity assessment framework, and Pekovitcs mega and major processes mapping. The data findings reveal a major variance between the perception and the requirements of digital transformation maturity. The examined institutions did not have a digital transformation vision, leadership, strategy, plan, champions, processes, controls, approach, communication, or proper return on investment. Many of the examined institutions viewed their maturity level at delivering and harvesting, when in fact they were at either designing or ambition.

In addition, digital transformation was more evident in enabling processes such as student administration services, library services, finance and accounting, but not as much in learning and teaching, research processes, and planning and governance processes.

The second research question is concerned with the digital transformation challenges. Leading challenges reported included challenges with holistic vision; personnel competency and IT skills; data structure, data processing, and data reporting; redundant systems; third party reporting systems; manual entries; and potential use by customers.

The challenges cited by the respondents in this study are not mutual exclusive; in fact, they are interrelated in multiple ways. While some of the challenges are more critical than others, the combination of those challenges create an environment that hinders digital transformation and business success by creating dependency, timeliness, integrity, availability, cost, efficiency, effectiveness, and integration issues.

VI. Recommendations

Digital transformation in education is inevitable. Higher education institutions should establish a clear vision, policies, strategies, and plans to support mature digital transformation. Institutions should communicate such policies, strategies, and vision, and receive feedback from internal and external customers and stakeholders about business process engineering and return on investment. Such plans should regularly be evaluated. Institutions should hire digital transformation experts in order to align the business strategy with digital transformation. Digital transformation should not be just another task handed down to IT personnel, or segmented among data owners. The difference between automation and digital transformation should be communicated, and training and awareness should be provided. A corner stone to all of this is to show solid management support to combat resistance to change and resistance to technology, and communicate the long term value gained from digital transformation. Digital transformation should be extended beyond the enabling processes to teaching and learning, governance, and research. Specifically, the areas of course, program, and student assessment and evaluation. The proposed framework of this study can be used as a scorecard to assess the digital transformation maturity in higher education, assisting institutions in pinpointing processes and criteria that require further attention.

References

[1] Al Tamimi & Company, report on "Digital Transformation in the Education Space: A Review of the Impact of New Technologies on Middle East Education", 2019.

[2] A. Bayler, and O. Oz, "Academicians' Views on Digital Transformation in Education. International Online Journal of Education and Teaching (IOJET) 2018, 5(4), 809-830.

[3] Bounfour, A. "From IT to Digital Transformation: A Long Term Perspective. Digital Future and Digital Transformation. 2020.

[4] D. Chaurasiya, "Digital Transformation: A Case Study", retrieved from https://www.researchgate.net/publication/340385414_Digital_Transformation?enrichId=rgreq-e65c791fef4fde3c103a4370994ad74e-XXX&enrichSource=Y292ZXJQYWdlOzM0MDM4NTQxNDtBUzo4NzU4ODQyMzE3OTA1OTNAMTU4NTgzODQ4MjI1MA%3D%3D&el=1_x_2&_esc=publicationCoverPdf, 2020.

[5] Deloitte Digital. The journey to government's digital transformation Report. 2019.

[6] M. H. Ismail, M. Khater, M. Zaki, "Digital Business Transformation and Strategy: What do we know so far?", 2017. Retrived from https://cambridgeservicealliance.eng.cam.ac.uk/resources/Downloads/Monthly%20Papers/2017 NovPaper_Mariam.pdf.

[7] Gartner, Inc.'s free "Creating Digital Value at Scale" webinar is a special report based on the opening keynote of the 2017 Gartner Symposium.

[8] E. Göll and J. Zwiers, “Technological Trends In The Mena Region: The Cases Of Digitalization And Information and Communications Technology (Ict), MENARA, 2018.

[9] Y. Limani, E. Hajrizi, L. Stapleton, M. Retkoceri, “Digital Transformation Readiness in Higher Education Institutions (HEI): The Case of Kosovo”, Science Direct, IFAC PapersOnLine 52-25 (2019).

[10] C. Mahlow, and A. Hediger, “Digital Transformation in Higher Education—Buzzword or Opportunity?”. Special Issue: Paradigm Shifts in Global Higher Education and eLearning. MAY 2019.

[11] V. Maltese, “Digital Transformation Challenges for Universities: Ensuring Information Consistency Across Digital Services”. Journal of Cataloging and Classification Quarterly. Volume 56, 2018 - Issue 7. 2018.

[12] A. Marks, M. Alali, K. Reitesema, “Learning Management Systems: a shift Toward Learning and Academic Analytics”. International Journal on Emerging Technologies in Learning. (2016).

[13] Digital McKinsey, “Digital Middle East: Transforming the region into a leading digital economy”. Mckinsey, 2016.

[14] NV. “Digital Transformation in Higher Education. [Online] Navitas Ventures. Available: https://www.navitasventures.com/wp-content/uploads/2017/08/HE-Digital-Transformation-Navitas_Ventures_-EN.pdf. 2017.

[15] A. Norton, s. Shroff, and N. Edwards, “Digital Trasnformation: An Enterprise Architecture Prespective”, Publish Nation Limited, UK, 2020.

[16] The Organisation for Economic Co-operation and Development OECD (2016). Digital Government Strategies for Transforming Public Services in the Welfare Areas.[Online] http://www.oecd.org/gov/digital-government/Digital-Government-Strategies-Welfare-Service.pdf.

[17] B. Parlak, “Dijital çağda eğitim: Olanaklar ve uygulamalar üzerine bir analiz [Education in Digital Age: An analysis on opportunities and practices], Süleyman Demirel University, Journal of Faculty of Economics and Administrative Sciences, 22(15), 1741-1759. 2017.

[18] I. Petkovics, P. Tumbas, P. Matkovic, Z. Baracskai, "Cloud Computing Support to University Business Processes in External Collaboration". Acta Polytechnica Hungarica, vol.11, no.3, pp.181-200, 2014.

[19] K. Sandkuhl, H. Lehmann, "Digital Transformation in Higher Education – The Role of Enterprise Architectures and Portals", Digital Enterprise Computing (DEC 2017), 2017.

[20] K. Schwab, K. Dördüncü sanayi devrimi [Fourth industrial revolution], İstanbul: Optimist Publications. 2016.

[21] K. Schwertner, “Digital transformation of business”. Trakia Journal of Science, 15(Suppl.1), pp.388-393. 2017.

[22] L. Seres, V. Pavlicevic, P. Tumbas, “Proceedings of INTED2018 Conference”. 5th-7th March 2018, Valencia, Spain.

[23] B. Solis, “8 Success Factors of Digital Transformation” Altimeter. Prophet Thinking. [online]: https://www.prophet.com/thinking/2016/02/brief-the-opposite-approach-8-success-factors-of-digital-transformation/.

[24] J. Thompson, “Books in the Digital Age”. New York, NY: John Wiley & Sons. 2013.

[25] A. Yıldırım, & H. Şimşek, “Sosyal bilimlerde nitel araştırma yöntemleri [Qualitative research methods in the social sciences”. Ankara: Seçkin Publishing.2013.

[26] R. Walker, J, Voce, J. & M, Jenkins, “Charting the development of technology-enhanced learning developments across the UK higher education sector: A longitudinal perspective” (2001–2012). Interactive Learning Environments, 24 (3), 438–455. doi:10.1080/10494820.2013.867888, 2016.

[27] J. Xiao, “Digital transformation in higher education: critiquing the five-year development plans (2016-2020) of 75 Chinese universities”, Distance Education, 40:4, 515-533, DOI: 10.1080/01587919.2019.1680272, 2019.

Temporal-based Optimization to Solve Data Sparsity in Collaborative Filtering

Ismail Ahmed Al-Qasem Al-Hadi[1]*, Mohammad Ahmed Alomari[2]
Eissa M. Alshari[3], Waheed Ali H. M. Ghanem[4], Safwan M Ghaleb[5]
Faculty of Ocean Engineering Technology and Informatics Universiti Malaysia Terengganu, Terengganu, Malaysia[1, 4, 5]
Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Terengganu, Malaysia[2]
Department of Data Science, Universiti Malaysia Kelantan, Kota Bharu 16100, Kelantan, Malaysia[3]

Abstract—**Collaborative Filtering (CF) is a widely used technique in recommendation systems. It provides personal recommendations for users based on their preferences. However, this technique suffers from the sparsity issue which occurs due to a high proportion of missing rating scores in a rating matrix. Several factorization approaches have been used to address the sparsity issue. Such techniques have also been considered to tackle other challenges such as the overfitted predicted scores. Nevertheless, they suffer from setbacks such as drift in user preferences and items' popularity decay. These challenges can be solved by prediction approaches that accurately learn the long-term and short-term preferences integrated with factorization features. Nonetheless, the current temporal-based factorization approaches do not accurately learn the convergence of the assigned k clusters due to a lower number of short-term periods. Additionally, the use of optimization algorithms in the learning process to reduce prediction errors is time-consuming which necessitates a faster optimization algorithm. To address these issues, a new temporal-based approach named TWOCF is proposed in this paper. TWOCF utilizes the elbow clustering method to define the optimal number of clusters for the temporal activities of both users and items. This approach deploys the whale optimization algorithm to accurately learn short-term preferences within other factorization and temporal features. Experimental results indicate that TWOCF exhibits a superior CF prediction accuracy achieved within a shorter execution time when compared to the benchmark approaches.**

Keywords*—*Collaborative filtering; matrix factorization; temporal-based approaches; whale optimization

I. Introduction

Nowadays, recommendation systems have become popular as they efficiently suggest items to customers according to their feedback and interests [1]. The major resources (data) used to create recommendations are customer profiles, item profiles, and user-item connections (i.e., customer scores to the suggested items) [2]. Collaborative filtering (CF), content-based filtering, demographic filtering, and hybrid filtering are four forms of filters employed in recommendation systems [3]. CF is one of the best prevalent recommendation techniques that provide users with personalized predictions based on their preferences. It relies solely on past users' rating scores on products and does not require the creation of explicit profiles. For example, CF utilizes the rating scores of neighbours to predict a list of items to the active user. However, CF suffers from three main issues: data sparsity [3], [4], cold start [5], [6], and scalability [7] [8]. This paper briefly discusses the approaches utilized to solve the sparsity issue.

Normally, CF-based recommendation systems arrange customers' product rating scores in form of a rating matrix. Customers rank a small number of products using rating scores. These scores are then arranged into a rating matrix. The rating matrix contains very few scores while others are unknown or sparse. This reduces the prediction accuracy of the CF technique. CF provides a list of recommendations to an active user based on his/her interests and according to the feedback of common users who rate some items that are rated by the active user. The feedback is calculated using the similarity and prediction assessments. The similarity assessment between common users and the active user will be infeasible or inconsistent if there is a higher proportion of missing rating scores in the rating matrix [9].

The optimization algorithms have proven successful in several areas such as healthcare [10], document processing [11], and recommendation systems [12]. Various factorization approaches have been used to solve the sparsity issue. These include an imputation-based matrix factorization [2], ensemble divide and conquer [13], and neighbourhood matrix factorization [14]. Although these approaches can learn factorization and latent features that influence the prediction accuracy of the CF technique [12], they cannot effectively learn temporal behaviors and temporal issues such as the drift in users' preferences and the popularity decay of items [15] [16].

The temporal collaboration model [17] merged factorization vectors, long-term preferences, and short-term preferences to enhance the efficiency of the CF technique. The short-term feedback is defined using the shrunk neighbors approach [18] [19] [17]. Furthermore, the short-term preferences are defined by a timestamp factor that determines time periods such as the number of years, seasons, months, and so on. This duration is used to assign *k* clusters to learn the short-term features using *k*-means clustering algorithm. The temporal-based approach [1] achieves higher prediction performance compared to other previous temporal approaches. However, this approach cannot be implemented on the small rating matrix since the *k*-clustering cannot be achieved (*k* problem).

*Corresponding Author

Also, bacteria foraging optimization algorithm has been used to learn the accurate temporal and factorization features [16], [15]. This algorithm provides different error values (fitness) in the search space. The error value increases in some iterations while in others, it decreases; thus, consuming time. To address all these challenges, a new temporal-based whale optimization approach named TWOCF is proposed in this paper.

The TWOCF approach uses an elbow clustering method [20] to accurately learn the precise number of clusters in the time matrix of users' activities. Additionally, TWOCF approach learns the accurate weights for short temporal features that are integrated with other factorization and long temporal features.

The unique aspects of this research can be summarized as follows:

- Solving the sparsity problem.
- Learning accurate latent features throughout the learning iterations.
- Solving the drift and decay issues.
- Reducing the running time of the learning process.

The factorization-based optimization approaches [15] [16] [1] have been focused on solving the temporal and sparsity issues based on personalization. These approaches have improved the accuracy of predicted rating scores while ignored the running time. Despite running time is a significant factor in online recommendation systems, most of the recommendation-based optimization approaches ignore running time. Therefore, TWOCF approach is proposed to solve temporal and sparsity issues with the purpose of improving the prediction accuracy and reducing running time.

The rest of the paper is constructed thus: Section 2 discusses the related works (including the factorization and temporal methods) and describes the whale optimization algorithm (WOA) whose methodology is explained in Section 3. In Section 4, the investigational findings are explained while Section 5 summarizes the main findings of this study.

II. Related Works

A. Matrix Factorization

Matrix factorization has been recently used for solving the sparsity and cold start issues associated with the CF technique in recommendation systems [21]. Matrix factorization is characterized by two features: baseline and latent features. The latent features are defined using the singular value decomposition algorithm [22]. Many factorization methods integrate latent and baseline features of users and items utilizing several formulae [7]. For example, the baseline formula can be used to predict the missing rating scores in the rating matrix as shown in Equation (1).

$$\widehat{r}_{uv} = \mu + B_u + B_v + p_u q_v^{\tau}, \tag{1}$$

where $\widehat{r}_{uv}$ is the expected value for the sparse score, μ is the global rate of all rating scores, B_u and B_v are the base values of users and items, respectively. The factors p_u and q_v^{τ} are the latent feedback of users and items, respectively. Additionally, the norm latent factor is used in several methods [23]. Equation (2) is another example where various latent features are integrated with the weight of γ to minimize the overfitting in the predicted rating scores.

$$\min \sum_{(u,v)\in k} (r_{uv} - p_u q_v^{\tau}) + \gamma(\|p_u\|^2 + \|q_v^{\tau}\|^2), \tag{2}$$

where $\|p_u\|$ and $\|q_v^{\tau}\|$ are the norm latent features of users and the norm of transport latent features of items, respectively.

Ensemble divide and conquer method is used to solve the rating scores' deviation that occurs when the rating scores are arranged in the memory. This method precisely learns latent features by rearranging the ratings in the rating matrix [13]. Nevertheless, the accuracy of this method is still low due to the drift in users' preferences and items' popularity decay. Thus, there is a need to study the positive effects of such temporal features for improving the prediction performance of recommendation systems [24]. Maintaining the Integrity of the Specifications.

B. Temporal Preferences

In recent years, temporal preferences and factorization factors are integrated within the collaborative-based approaches to solve sparsity problems [25]. The temporal dynamics method defines the time features by splitting the timeline into constant numbers of bins [25] while the user preferences are altered over time. This approach minimizes the overfitting of the predicted rating scores in the optimization latent space by a global weight (which is characterized by weakness in terms of personality). Generally, temporal preferences are long or short-termed [17].

- Long-term preferences

The long-term approach [17] manages the time of recorded scores to calculate the long-term preferences as expressed in Equation (3).

$$B_{uv} = \mu + B_u \frac{e - t_{uv}}{e - s} + B_u \frac{t_{uv} - s}{e - s} + B_v, \tag{3}$$

where s and e represent the first and last time preferences that are sequentially recorded in the rating matrix, t_{uv} is the current time item v is rated by user u. The long term preferences of users are defined in Equation (4) [15].

$$\vartheta_u = \exp\left[\frac{-(t_e^u - t_s^u)}{t_e^u}\right], \tag{4}$$

where ϑ_u is the temporal weight of user u, t_s^u is the first time and t_e^u is the last time that user u gave ranking scores.

Similarly, the long-term preferences of items are defined in Equation (5) [15].

$$\vartheta_v = \exp\left[\frac{-(t_e^v - t_s^v)}{t_e^v}\right], \tag{5}$$

where ϑ_v is the temporal weight of item v, t_s^v is the first time and t_e^v is the last time that item v was rated by users.

- Short-term preferences

The short-term based latent model learns the drift of users' preferences by incorporating the factorization features with the neighbors' latent feedback throughout a session, e.g., one month, one season, one year, etc. The temporal interaction model [17] combines the preferences of both short-term and long-term to address the drift in users' preferences. Nevertheless, both the short-term model and the temporal integration model [17] have limitations in terms of discovering the drift and time decay.

The short-term based factorization model [16] utilizes the *k*-means algorithm to divide the time matrix into a number of clusters based on the number of short-term periods such as the number of months. This model also deploy a bacterial foraging optimization algorithm to learn the best short-term weight to be integrated within the factorization features during the iterative learning procedure [16]. However, the sparse timestamp matrix of some active users cannot be divided into a certain number of clusters if the number of common users who rated the active user's item of interest is not sufficient (*k* problem).

In addition, longer execution time is required to learn the accurate temporal and factorization features, especially if the number of common users in the rating matrix is large. Hence, to address these issues while improving on our earlier works [1] [16], a new temporal-based approach named TWOCF is proposed. TWOCF approach assigns optimum temporal preferences and accurately learns the factorization and latent features using WOA.

C. Whale Optimization Algorithm

Whales are considered the largest animals in the world. They are always awake, quite smart, and sensitive to feelings. A truly exciting optimization idea is taken from the humpback whales, a group of whales with a unique searching technique named the bubble-net feeding method. The humpback whales only use the bubble-net for feeding. They would rather chase a bunch of krill or little fishes close to the surface. This hunting is completed by producing distinctive bubbles along a circle [26]. The feeding behaviour of humpback whales is mathematically represented by WOA to solve optimization problems.

III. Methodology

A. MovieLens Dataset

Several prediction approaches have utilized the MovieLens dataset to evaluate the performances of recommendation systems [15], [27]. All records in this dataset have been collected using the MovieLens website (movielens.umn.edu) throughout seven months from September 1997 to April 1998. There are 100,000 rating scores collected by 943 customers for 1682 movies. Each rating score that is collected by users is saved with its timestamp info. In the data set, each user has rated at least 20 movies, rating range is 1-5, each user can rate the movie as 1, 2, 3, 4, 5. The higher the user's rating of the movie, the more interested the user is in the movie. The sparseness of dataset can be calculated as: 1 – (100000/ (943 ∗ 1682)) = 0.936953. Additional features are collected by users such as age, gender, occupation, etc. Further features for items are movie id, movie title, release date, video release date, and genre. Similar to the benchmark methods, three significant features are considered in this paper. They are rating score, timestamp, and item genre.

B. Evaluation and Benchmark Methods

This work focuses on resolving the sparsity, overfitting, drift, and decay issues to improve on the prediction performance of the CF technique. The root mean square error (RMSE) function is employed for performance evaluation. RMSE has been utilized in many prediction approaches [15], [28] to evaluate the performance of the CF technique. A lower RMSE value indicates a higher prediction accuracy.

The experimental model can be benchmarked with some approaches implemented on the same dataset. These approaches are the CF, Ensemble Divide and Conquer [13], Temporal Dynamics [25], Long Temporal-based Factorization [15], Short Temporal-based Factorization [16], and Temporal-based Factorization Approach [1]. These approaches are implemented using MovieLens dataset underscoring [1-5]. Additionally, Short Temporal [16], and Temporal-based Factorization Approach [1] are implemented for three short terms which are 1 month, 2 weeks, and 1 week. The dataset contains 7 periods of 1 month, 15 periods of 2 weeks, and 30 periods of 1 week.

C. Experimental TWOCF Approach

The TWOCF method incorporates the long-term and short-term preferences with the factorization features to furnish the sparse rating matrix in order to yield accurate predictions. Besides the sparsity issue, three other challenges will be addressed by TWOCF. These are overfitting, drift, and decay as illustrated in subsequent subsections.

1) Assigning the temporal preferences: There are two kinds of temporal preferences: short-term and long-term preferences. The short temporal-based factorization model [16] analyzes the time matrix using the k-means algorithm. Similarly, the number of clusters is assigned according to the total number of sessions (e.g., MovieLens dataset contains 7 sessions and each session spans one month). However, this technique is suitable for smaller k periods (such as number of years or seasons) and not larger values (such as the number of weeks). This is the case for most users when the number of clusters formed is less than the target k after several convergences.

In addition, there are other user-centric differences. For example, some users are active only within a very short time while others are active for a longer time. These make numbering clusters in periods inaccurate. Thus, the best way to solve these challenges is by using clustering algorithms that can define the number of clusters accurately. The elbow clustering method is one of the most common methods used to determine the optimal values of clusters [29], [30]. In this work, the elbow clustering method is used to tackle the challenge of determining the number of clusters in the sparse timestamp matrix.

Fig. 1 shows a simple formation of the whale members using the elbow clustering method. Sometimes, the learning process stops because the clusters created by the k-means method are less than the required number of clusters. Equations (4) and (5) [15] are used to assign long-term preferences for users and items, respectively.

2) Integrating the temporal preferences with factorization features: This experimental work is intended to (i) predict the missing rating scores in the rating matrix following the CF technique and (ii) address other limitations such as overfitted predicted scores, drift in the users' preferences and decay in the popularity of items. The factorization and latent features that are integrated with temporal preferences are learned as follows:

$$\Re_{uv} = \mu + T_{\Omega_x^u} B_u \vartheta_u + T_{\varphi_y^v} B_v \vartheta_v + p_u q_v^\tau, \qquad (6)$$

where ϑ_u and ϑ_v are the long preference of user *u* and item *v*, respectively. $T_{\Omega_x^u}$ and $T_{\varphi_y^v}$ are short-term weights assigned by cluster number *x* and cluster number *y* for user *u and* item *v*, respectively. B_u and B_v are the baseline value of user *u* and item *v* respectively. p_u is the latent feedback of user *u*, and q_v^τ is the transport latent feature of item *v*.

The WOA is used to update the short-term weights of users and items. These weights will be integrated with the factorization features and with the long-term preferences using Equation (6) to reduce the overfitted predicted rating scores. Subsequently, the WOA learns the drift in the preferences of users and the decay in the popularity of items. This helps to reduce the effects of these negative factors throughout the iterative learning process using Equations (7) and (8), respectively.

$$\Re_u = T_{\Omega_x^u} \left[B_u \vartheta_u + \sum_{u=1}^{M} (p_u)^2 \right], \qquad (7)$$

$$\Re_v = T_{\varphi_y^v} \left[B_v \vartheta_v + \sum_{v=1}^{N} (q_v^\tau)^2 \right], \qquad (8)$$

where *M* is the number of common users who provide rating scores for items and *N* is the number of all items rated by the active user. To obtain the best performance, this work integrates the latent feedback of Equation (6) with Equations (7) and (8) using Equation (9).

$$\widehat{\Re}_{uv} = \Re_{uv} + \Re_u + \Re_v, \qquad (9)$$

where $\widehat{\Re}_{uv}$ is the predicted value for the missing rating score value by user *u* for item *v*. In Equation (9), all factors are computed by TWOCF in the first iteration and these values cannot change in other iterations except the short duration factors that can be updated throughout the optimization stages using the WOA. The predicted rating score $\widehat{\Re}_{uv}$ value can be updated throughout the iteration loops using the short-term factors $T_{\Omega_x^u}$ and $T_{\varphi_y^v}$ which provide a positive effect in improving the accuracy prediction of CF technique by reducing the error value.

3) Integrating WOA within TWOCF approach: WOA is integrated with TWOCF approach to optimize the prediction of the CF technique. This is aimed at addressing sparsity, overfitting, drift, and decay issues. TWOCF updates the weights of short temporal preferences throughout the iterative learning process managed by WOA. Three feeding behaviors of the humpback whales are briefly discussed as follows:

- Encircling prey

Humpback whales identify the locations of small fishes, then engulf them [31]. WOA algorithm assumes that the recent superlative candidate result is the objective prey or is near to the optimum. After the finest search agent is identified, the other search representatives will later attempt to revise their positions for finding the best search representative. This performance is characterized by Equations (10) and (11).

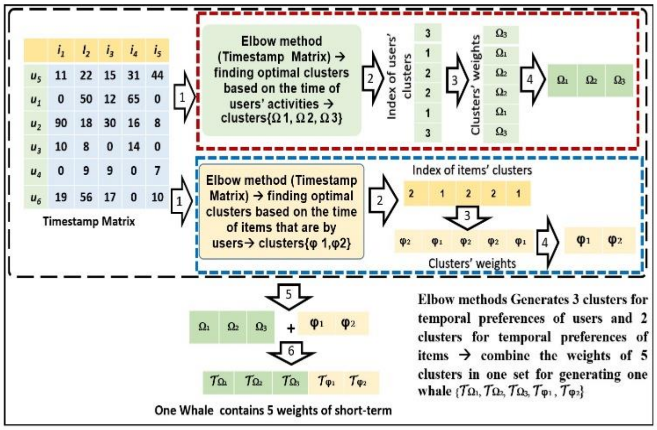
$$\vec{D} = \left| \vec{C}.\vec{T}^*(i) - \vec{T}(i) \right|, \qquad (10)$$

$$\vec{T}(i+1) = \vec{T}^*(i) - \vec{A}.\vec{D}, \qquad (11)$$

where *i* is the number of the current iteration, $\vec{A}$ and $\vec{C}$ are coefficient vectors. In this work, $\vec{T}$ represents the short duration weights which is the position vector of the prey. $\vec{T}^*$ is the position vector of the best solution (accurate short duration weights) obtained so far. The vectors $\vec{A}$ and $\vec{C}$ are computed by Equations (12) and (13) as follows:

$$\vec{A} = 2\vec{\varepsilon}.\vec{r} - \vec{\varepsilon}\ , \qquad (12)$$

$$\vec{D} = 2.\vec{r}\ , \qquad (13)$$



	i_1	i_2	i_3	i_4	i_5
u_5	11	22	15	31	44
u_1	0	50	12	65	0
u_2	90	18	30	16	8
u_3	10	8	0	14	0
u_4	0	9	9	0	7
u_6	19	56	17	0	10

Timestamp Matrix

Elbow method (Timestamp Matrix) → finding optimal clusters based on the time of users' activities → clusters{Ω 1, Ω 2, Ω 3}

Elbow method (Timestamp Matrix) → finding optimal clusters based on the time of items that are by users→ clusters{φ 1,φ2}

Elbow methods Generates 3 clusters for temporal preferences of users and 2 clusters for temporal preferences of items → combine the weights of 5 clusters in one set for generating one whale {TΩ1, TΩ2, TΩ3, Tφ1 , Tφ2}

One Whale contains 5 weights of short-term

Fig. 1. A Simple Example of Forming a Whale's Members.

where $\vec{\varepsilon}$ is linearly reduced from 2 to 0 throughout the iterative process and $\vec{r}$ is a random vector from 0 to 1. Equation (11) allows each search agent to update the temporal weight that is close to the current best position and replicate circling the prey.

- Bubble-net attacking method

WOA mathematical modeling involves two phases modeled as follows:

1) Shrinking encircling mechanism: In Equation (12), the value of $\vec{\varepsilon}$ is reduced, then $\vec{A}$ is also reduced by $\vec{\varepsilon}$. Therefore, the new position of a search agent can be defined between the current temporal positions and the best temporal positions by setting random values for $\vec{A}$ from 0 to 1.

2) Spiral updating position: The distance between the whale and its prey can be calculated using the weights of temporal features and the best weights of temporal features. A spiral equation is then formed between the position of whale and prey as shown in Equation (14).

$$\vec{T}(i+1) = \vec{D}'.\exp(b\omega).\cos(2\pi\omega) + \vec{T}^{*}(i), \tag{14}$$

where $\vec{D}' = \left|\vec{T}^{*}(i) - \vec{T}(i)\right|$ denotes the distance between the whale and its prey, *b* is a constant number to define the form of logarithmic rise, and ω is a random number in [-1, 1]. The humpback whales swim across the prey contained by a disappearing circle along a spiral-shaped path concurrently. This simultaneous behavior takes a 50% probability to choose either the shrinking surrounding structure (Equation (6)) or the spiral model for updating the temporal weights (Equation (9)) during the optimization as shown below.

$$\begin{aligned}
&if\,(p < 0.5)\\
&\vec{T}(i+1) = \vec{D}'.\exp(b\omega).\cos(2\pi\omega) + \vec{T}^{*}(i)\\
&elseif\,(p >= 0.5)\\
&\vec{T}(i+1) = \vec{T}^{*}(i) - \vec{A}.\vec{D}
\end{aligned}$$

where *p* is a random number in [0,1].

- Search for prey (Exploration phase)

In this situation, $\vec{A}$ is utilized by the random values greater than 1 or lesser than −1 for forcing the search agent to change the current whale's location [31]. This allows the WOA algorithm to perform a global search that is modeled using Equations (15) and (16).

$$\vec{D} = \left|\vec{C}.\vec{T}_{rand} - \vec{T}\right|, \tag{15}$$

$$\vec{T}(i+1) = \vec{T}_{rand} - \vec{A}.\vec{D}, \tag{16}$$

where $\vec{T}_{rand}$ is a random temporal vector. The phases of TWOCF approach are further detailed in the TWOCF algorithm.

TWOCF algorithm

```
Procedure of Factorization:
    Input: Rating Matrix
    Output: $\mu, B_u, B_v, p_u, q_v, q_v^{\tau}$
        $B_u = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(r_{uv} - \mu)^2}$
        $B_v = \sqrt{\frac{1}{m}\sum_{v=1}^{m}(r_{uv} - \mu)^2}$
        $[p_u, d, q_v] = SVD(RatingMatrix)$
        $q_v^{\tau}$ =transpose ($q_v$)
Procedure of Long-term:
    Input: Timestamp Matrix
    Output: $\vartheta_u, \vartheta_v$
        $\vartheta_u = \exp\left[-(t_e^u - t_s^u)/t_e^u\right]$
        $\vartheta_v = \exp\left[-(t_e^v - t_s^v)/t_e^v\right]$
Procedure of Short-Term:
    Input: Timestamp Matrix
    Output: $[T_{\Omega}, T_{\Omega_2}, T_{\Omega_3}, \ldots\ldots, T_{\Omega_x}, T_{\varphi_1}, T_{\varphi_1}, T_{\varphi_1}, \ldots\ldots., T_{\varphi_y}]$
        Clustering users
            [Index^u, x] = Elbow clustering Method (Timestamp Matrix)
            $T_{\Omega_1}, T_{\Omega_2}, T_{\Omega_3}, \ldots\ldots, T_{\Omega_x}$ = [Index^u, x]
        Clustering items
            [Index^v, y] = Elbow clustering Method (Timestamp Matrix)
            $T_{\varphi_1}, T_{\varphi_1}, T_{\varphi_1}, \ldots\ldots., T_{\varphi_y}$ = [Index^v, y]
        Managed weights of users and items in one whale members
            weights=
            $[T_{\Omega}, T_{\Omega_2}, T_{\Omega_3}, \ldots\ldots, T_{\Omega_x}, T_{\varphi_1}, T_{\varphi_1}, T_{\varphi_1}, \ldots\ldots., T_{\varphi_y}]$
Procedure of Collaborative Filtering (Fitness):
    Input: Rating Matrix, Timestamp Matrix
    Output: RMSE value
        Call Procedure of Factorization
        $\mu, B_u, B_v, p_u, q_v, q_v^{\tau}$
        Call Procedure of Long-term to obtain $\vartheta_u, \vartheta_v$
        Call Procedure of WOA to update
        $[T_{\Omega}, T_{\Omega_2}, T_{\Omega_3}, \ldots\ldots, T_{\Omega_x}, T_{\varphi_1}, T_{\varphi_1}, T_{\varphi_1}, \ldots\ldots., T_{\varphi_y}]$
        Predicting the missing values in Rating Matrix
            $\Re_{uv} = \mu + T_{\Omega_x^u} B_u \vartheta_u + T_{\varphi_y^v} B_v \vartheta_v + p_u q_v^{\tau}$
            $\Re_u = T_{\Omega_x^u}\left[B_u \vartheta_u + \sum_{u=1}^{M}(p_u)^2\right]$
            $\Re_v = T_{\varphi_y^v}\left[B_v \vartheta_v + \sum_{v=1}^{N}(q_v^{\tau})^2\right]$
            $\widehat{\Re}_{uv} = \Re_{uv} + \Re_u + \Re_v$
        Getting similarity Values using Cosine Function
            $sim(u_a, u_c) = \sum_{v=1}^{M}(r_{u_a,item_v}.r_{u_c,item_v}) \Big/ \left[\sqrt{\sum_{v=1}^{M} r_{u_a,item_v}^2}.\sqrt{\sum_{v=1}^{M} r_{u_c,item_v}^2}\right]$
        Getting predicted Values for the items rated by active user
        $p_v = \mu_{u_a} + \left[ \sum_{c=1}^{M} sim(u_a,u_c)(r_{u_c,item_v} - \mu_{u_c}) \right] \Big/ \sqrt{\sum_{c=1}^{M} |sim(u_a,u_c)|}$
          (p_v is real/infinite value in [0-5])
        Getting fitness Value by RMSE function
        $RMSE_{u_a} = \sqrt{\frac{1}{N} \sum_{v=1}^{N} (r_v - p_v)^2}$

Procedure of WOA:
    Input: $[T_{\Omega},T_{\Omega_2},T_{\Omega_3},.....,T_{\Omega_x},T_{\varphi_1},T_{\varphi_1},T_{\varphi_1},.......,T_{\varphi_y}]$
    Output: updated $[T_{\Omega},T_{\Omega_2},T_{\Omega_3},.....,T_{\Omega_x},T_{\varphi_1},T_{\varphi_1},T_{\varphi_1},.......,T_{\varphi_y}]$
        Assign the parameters of WOA
        $\vec{\varepsilon} = random(between 2 \& \&0)$
        $\vec{r} = random(between 0 \& \&1)$
        $\vec{A} = 2\vec{\varepsilon}.\vec{r} - \vec{\varepsilon}$
        $\vec{D} = 2.\vec{r}$
        Search agent (whale)=10
        T*=the best temporal weights (the best search agent)
        While iteration <MaxIteration
            For each search agent
            $[T_{\Omega},T_{\Omega_2},T_{\Omega_3},.....,T_{\Omega_x},T_{\varphi_1},T_{\varphi_1},T_{\varphi_1},.......,T_{\varphi_y}]$
            $\vec{D} = |\vec{C}.\vec{T}_{rand} - \vec{T}|$
            Create a random number in [0-1] → p
            $if\,(p < 0.5)$
            $\vec{T}(i+1) = \vec{D}'.\exp(b\omega).\cos(2\pi\omega) + \vec{T}^*(i)$
            $elseif\,(p >= 0.5)$
            $\vec{T}(i+1) = \vec{T}^*(i) - \vec{A}.\vec{D}$
            Call Procedure of Collaborative Filtering (Fitness) → $RMSE_{u_a}$
            if RMSE<fitness
             fitness=RMSE
             update $[T_{\Omega},T_{\Omega_2},T_{\Omega_3},....,T_{\Omega_x},T_{\varphi_1},T_{\varphi_1},T_{\varphi_1},....,T_{\varphi_y}]$
            end if
        End While
Main TWOCF
    Data preparation based on personalized
        Assign the active user $u_a$
        Arrange scores value in Rating Matrix
        Arrange timestamp value in Timestamp Matrix
            Active user is arranged in the first row and the items rated by active user are arranged as columns
            Add any user provides scores to any item rated by active user to the Matrix
             → Rating Matrix, Timestamp Matrix
    Procedure of Short-Term
    $[T_{\Omega},T_{\Omega_2},T_{\Omega_3},.....,T_{\Omega_x},T_{\varphi_1},T_{\varphi_1},T_{\varphi_1},.......,T_{\varphi_y}]$
        Assign WOA vectors
            Search Agents= 10
            dim=x+y → (number of short-term weights)
            The upper bound of each weight in dim = 1
            The lower bound of each weight in dim = -1
            MaxIteration=300
        Initialize random values for whale members →
         Matrix (Search Agents, dim) =rand
        Repeat
            Call Procedure of WOA → RMSE
                If fitness<RMSE
                    Fitness=RMSE
                    Update $[T_{\Omega},T_{\Omega_2},T_{\Omega_3},.....,T_{\Omega_x},T_{\varphi_1},T_{\varphi_1},T_{\varphi_1},.......,T_{\varphi_y}]$
        Until complete Max iteration
    Output:
    The accurate Rating Matrix with predicted missing scores
    The accurate list of recommendations for the active user
```

IV. Experimental Results

A. The Effect of TWOCF in Solving the Overfitting, Drift, and Decay

The TWOCF approach is proposed to address the weaknesses of factorization and temporal-based factorization approaches. Experimental results show that the TWOCF approach exhibits significantly superior performance with respect to learning the accurate temporal and factorization features by reducing overfitted predicted scores, tracing decay in the popularity of items, and tracing drifting in the users' preferences. Table I presents the experimental results obtained after implementing the TWOCF approach on MovieLens dataset under-scoring [1-5].

The first column in Table I contains 31 active users. The second and third columns indicate the dimensions of the rating matrix for assigning the learning search space. Columns 4 and 5 show the cluster's number based on users' and items' dimensions, respectively. The different numbers of clusters in each matrix refer to personality behaviors of users. It is worthy of note that this cannot be learned accurately using a specific number of clusters. The sixth column shows several numbers of whale members by which the TWOCF approach accurately learns the features of users and items.

In column 7, the execution time of learning procedures varies according to the dimension space of each matrix. The shortest execution time is 11 seconds while the longest execution time is 487 seconds. The last column indicates the prediction accuracy of the CF technique according to RMSE values. Here, a lower value indicates a higher prediction accuracy. Using the TWOCF approach, results ranging from 0.523 to 0.997 with an average of 0.764 are obtained.

The learning processes by the TWOCF approach are visualized in Fig. 2 to show its ability to reduce the RMSE values throughout the iteration loops. Fig. 2 shows the effectiveness of the TWOCF approach in accurately learning the behaviours of users and items throughout the learning iteration. This improves the CF technique's speed and learning accuracy.

B. Comparison of the Performances of CF, Factorization, and Factorization-based Temporal Approaches

Here, the TWOCF approach is evaluated by comparing its effectiveness in reducing the RMSE value with other benchmark approaches described in Section 2. The TWOCF and benchmark approaches compared are implemented using one Test-Set (contains 31 rating matrices) to predict the missing scores in each rating matrix. In addition, the contributions of the tested approaches to solve the issues that are reviewed in this article are summarized in Table II.

TABLE I. THE EXPERIMENTAL RESULTS OF THE TWOCF APPROACH

Active User	**Matrix Dimensions**		**Clusters numbers**		**Whale members**	**Execution Time (Seconds)**	**RMSE value**
	Common Users	**Items**	**Common Users**	**Items**			
u1	937	43	24	7	31	21	0.994
u2	928	207	21	14	35	124	0.723
u3	943	299	21	15	36	205	0.784
u4	920	26	24	6	30	13	0.979
u5	916	32	25	6	31	15	0.541
u6	883	63	21	7	28	26	0.693
u7	939	132	22	11	33	64	0.740
u8	871	24	24	5	29	11	0.670
u9	943	138	23	11	34	70	0.802
u10	876	20	26	5	31	10	0.885
u11	914	147	22	12	34	84	0.872
u12	942	102	23	11	34	50	0.665
u13	931	31	24	6	30	17	0.628
u14	930	36	24	6	30	18	0.555
u15	943	539	23	20	43	487	0.678
u16	938	60	22	8	30	27	0.829
u17	884	30	24	6	30	14	0.680
u18	943	63	24	8	32	29	0.738
u19	875	22	23	5	28	11	0.845
u20	840	89	18	10	28	37	0.826
u21	940	108	21	9	30	52	0.890
u22	941	224	22	11	33	148	0.944
u23	885	115	24	11	35	52	0.849
u24	845	30	26	6	32	13	0.882
u25	895	33	26	6	32	16	0.731
u26	938	55	23	8	31	26	0.523
u27	890	26	25	6	31	13	0.638
u28	933	196	23	13	36	124	0.722
u29	943	269	23	16	39	184	0.679
u30	911	45	23	7	30	20	0.713
u31	941	63	23	8	31	29	0.997
Average	915	105	23	9	32	65	0.764

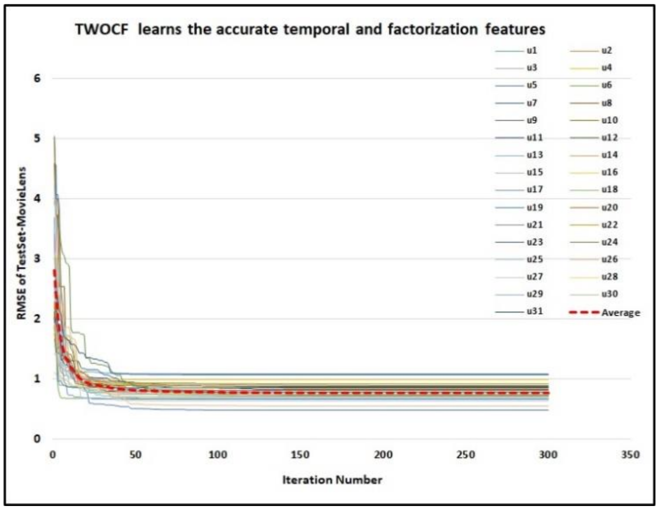


Fig. 2. TWOCF Improves the Prediction Accuracy of the CF Technique.

The CF technique has provided the lowest accuracy prediction because of the negative effects of sparsity, drift, and decay issues. The highest RMSE value represents the lowest accuracy prediction. Ensemble Divide and Conquer [13] is used to solve the sparsity and missing the accurate location of data when arrange this data into memory. The output results show better performance compared to the CF. However, this method has a weakness in terms of drift and decay. Long Temporal based Factorization [15] is used to solve the sparsity, overfitting and decay issues by learning the long-term features through the convergence of genres features of the items.

TABLE II. THE CF PERFORMANCE USING THE PREDICTION APPROACHES

Approach	Objectives				Long duration	Short duration assigned by	Execution time (Seconds)	Average RMSE
	Sparsity	Overfitting	Drift	Decay				
CF								0.957
Ensemble Divide and Conquer [13]	√							0.954
Temporal Dynamics [25]	√	√	√	√	√	Time slices		0.951
Long Temporal based Factorization [15]	√	√		√	√			0.947
Short Temporal based Factorization [16]	√	√	√			1 week	4256	0.851
						2 weeks	4049	0.863
						1 month	3921	0.870
Temporal based Factorization [1]	√	√	√	√	√	1 week	1316	**0.818**
						2 weeks	4128	0.819
						1 month	4197	0.843
TWOCF	√	√	√	√	√	Auto clustering	**65**	**0.764**

The short temporal features (column 7) are defined by different duration factors to achieve accurate solutions. For example, Temporal Dynamics [25] defines the short-term by time slices. However, it has weaknesses in terms of personality.

Short Temporal-based Factorization [16] and Temporal-based Factorization [1] approaches defined the short-term periods using the k-means algorithm where the number of the clusters is assigned based on the number of certain times (e.g., in MovieLens the whole time of users' activities can be divided into 7, 15, or 30 clusters when assigning one month, 2 weeks, or one week period, respectively).

The Short Temporal-based Factorization [16] is used to solve drift of users and ignored the issue of decay during long duration which reduce the accuracy prediction performance of the CF. Temporal-based Factorization [1] approach is used to solve all issues. Its result is the best comparing to the benchmark approaches. However, the benchmark approaches are ignored the results of running time due to the iteration of optimization procedure is very slow as shown in Table II (e.g. minimum average of running time is 1316 second). Bacterial foraging optimization algorithm is used with the last three benchmark approaches of temporal and its experimental running times are slow as shown in Table II which represent as a significant weakness. The recommendation systems need high accuracy as well as faster running time. As observed in Table II, the studied approaches have different executing time and accuracy. It is obvious that approaches with high accuracy have long run time, e.g., Temporal-based Factorization [1] provides lower RMSE (high accuracy) but with a long executing time.

Distinctively, TWOCF approach learns the accurate features of each user within the smallest execution time. Additionally, the TWOCF approach provides the highest accuracy prediction compared to the other benchmark approaches. This means that the TWOCF approach has the best performance and can deal with all kinds of matrices as the number of clusters is assigned automatically. Moreover, TWOCF approach performs best in reducing the overfitted predicted scores and accurately learning the temporal features throughout the learning iteration, which reducing the negative effect of drift and decay in the prediction performance of the CF technique.

V. Conclusion

Recommendation systems are becoming popular because they can efficiently recommend products to customers based on their interests. CF-based recommendation systems perform well since they consider the rating matrix in their execution. Nevertheless, CF suffers from the sparsity issue which is usually tackled using factorization approaches. Similarly, overfitting is another challenge mainly addressed using optimization approaches. Additionally, the drift in users' preferences and items' popularity decay addressed by Temporal-based factorization approaches are also major setbacks. Although the current solutions achieve some level of accuracy, there is still room for improvement. For example, dividing the temporal activities throughout the duration search space, reducing the runtime of the execution process, and lowering the error values of the predicted rating scores.

The TWOCF approach is proposed to render timely and accurate predictions within the rating matrix by accurately learning users' preferences and items' popularity pattern. TWOCF adopts the elbow clustering method to obtain the optimal number of temporal clusters. Also, the short-term weights of generated clusters are integrated with the factorization features for predicting the missing scores in the rating matrix. Results show that the TWOCF approach outperforms the benchmark schemes, improves the accuracy of the CF technique, and reduces its execution time.

References

[1] A. A.-Q. Al-Hadi, N. Mohd Sharef, S. M. Nasir, and M. Norwati, "Temporal based factorization approach for solving drift and decay in sparse scoring matrix," Int. Conf. Soft Comput. Data Mining, Springer, vol. 700, pp. 340–350, 2018.

[2] M. Ranjbar, P. Moradi, M. Azami, and M. Jalili, "An imputation-based matrix factorization method for improving accuracy of collaborative filtering systems," Eng. Appl. Artif. Intell., vol. 46, pp. 58–66, 2015.

[3] M. H. Mohamed, M. H. Khafagy, and M. H. Ibrahim, "Recommender Systems Challenges and Solutions Survey," Int. Conf. Innov. Trends Comput. Eng., pp. 149–155, 2019.

[4] D. Chae, J. Kim, and S. Kim, "AR-CF : Augmenting Virtual Users and Items in Collaborative Filtering for Addressing Cold-Start Problems," Sigir 2020, no. July, pp. 1251–1260, 2020.

[5] Z. Zhang, Y. Kudo, T. Murai, and Y. Ren, "Improved covering-based collaborative filtering for new users' personalized recommendations," Knowl. Inf. Syst., vol. 62, no. 8, pp. 3133–3154, 2020.

[6] B. Alhijawi and Y. Kilani, "The recommender system: A survey," Int. J. Adv. Intell. Paradig., vol. 15, no. 3, pp. 229–251, 2020.

[7] A. Pujahari and D. S. Sisodia, "Pair-wise Preference Relation based Probabilistic Matrix Factorization for Collaborative Filtering in Recommender System," Knowledge-Based Syst., vol. 196, no. xxxx, p. 105798, 2020.

[8] M. Nilashi et al., "Preference learning for eco-friendly hotels recommendation: A multi-criteria collaborative filtering approach," J. Clean. Prod., vol. 215, pp. 767–783, 2019.

[9] I. A. A. Q. Al-Hadi, N. M. Sharef, M. N. Sulaiman, and N. Mustapha, "Review of the temporal recommendation system with matrix factorization," Int. J. Innov. Comput. Inf. Control, vol. 13, no. 5, pp. 1579–1594, 2017.

[10] N. Zainal, I. A. A. Q. Al-Hadi, S. M. Ghaleb, H. Hussain, W. Ismail, and A. Y. Aldailamy, "Predicting MIRA Patients' Performance Using Virtual Rehabilitation Programme by Decision Tree Modelling," Recent Adv. Intell. Syst. smart Appl. Cham Springer, pp. 451–462, 2020, doi: 10.1007/978-3-030-47411-9_24.

[11] A. Al-badarneh, M. Ali, and S. M. Ghaleb, "An Improved Classifier for Arabic," J. Converg. Inf. Technol., vol. 11, no. 3, pp. 69–84, 2016.

[12] I. A. A. Al-hadi, N. M. Sharef, N. Mustapha, and M. Nilashi, "Latent based temporal optimization approach for improving the performance of collaborative filtering," PeerJ Comput. Sci., pp. 1–25, 2020, doi: 10.7717/peerj-cs.331.

[13] I. A. A. Q. Al-Hadi, N. M. Sharef, M. N. Sulaiman, and N. Mustapha, "Ensemble divide and conquer approach to solve the rating scores' deviation in recommendation system," J. Comput. Sci., vol. 12, no. 6, pp. 265–275, 2016.

[14] M. Guo, J. Sun, and X. Meng, "A Neighborhood-based Matrix Factorization Technique for Recommendation," Ann. Data Sci., vol. 2, no. 3, pp. 301–316, 2015.

[15] I. A. A.-Q. Al-Hadi, N. M. Sharef, M. N. Sulaiman, and N. Mustapha, "Temporal-Based Approach to Solve Item Decay Problem in Recommendation System," Adv. Sci. Lett., vol. 24, no. 2, pp. 1421–1426, 2018.

[16] I. A. A.-Q. Al-Hadi, N. M. Sharef, M. N. Sulaiman, and N. Mustapha, “Bacterial foraging optimization algorithm with temporal features to solve data sparsity in recommendation system,” ICC ’17 Proc. Second Int. Conf. Internet things, Data Cloud Comput. ACM., pp. 148:1--148:6, 2017.

[17] F. Ye and J. Eskenazi, “Feature-Based Matrix Factorization via Long- and Short-Term Interaction,” Knowl. Eng. Manag., pp. 473–484, 2014.

[18] Y. Koren, “Factorization meets the neighborhood: A multifaceted collaborative filtering model,” Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 426–434, 2008.

[19] D. Yang, T. Chen, W. Zhang, and Y. Yu, “Collaborative filtering with short term preferences mining,” Proc. 35th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR ’12, no. 2, p. 1043, 2012.

[20] “Sebastien De Landtsheer (2020). kmeans_opt (https://www.mathworks.com/matlabcentral/fileexchange/65823-kmeans_opt),” MATLAB Cent. File Exch. Retrieved Sept. 3, 2020.

[21] R. Lara-Cabrera, Á. González-Prieto, angel gonzalez prieto@upm Es, F. Ortega, and jesus bobadilla@upm es Jesús Bobadilla, “Evolving matrix-factorization-based collaborative filtering using genetic programming,” Appl. Sci., vol. 10, no. 2, 2020.

[22] M. Nilashi, O. Ibrahim, and K. Bagherifard, “A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques,” Expert Syst. Appl., vol. 92, pp. 507–520, 2018.

[23] X. Liu, J. Zhang, and C. Yan, “Towards context-aware collaborative filtering by learning context-aware latent representations,” Knowledge-Based Syst., vol. 199, pp. 1–13, 2020.

[24] J. Vinagre, A. M. Jorge, and J. Gama, “An overview on the exploitation of time in collaborative filtering,” Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 5, no. 5, pp. 195–215, 2015.

[25] Y. Koren, “Collaborative filtering with temporal dynamics,” Commun. ACM, vol. 53, no. 4, pp. 89–97, 2010.

[26] S. Adhirai, R. P. Mahapatra, and P. Singh, “The Whale Optimization Algorithm and Its Implementation in MATLAB,” Int. J. Comput. Inf. Eng., vol. 12, no. 10, pp. 815–822, 2018.

[27] H. Koohi and K. Kiani, “A new method to find neighbor users that improves the performance of Collaborative Filtering,” Expert Syst. Appl., vol. 83, pp. 30–39, 2017.

[28] F. Pajuelo-Holguera, J. A. Gómez-Pulido, F. Ortega, and J. M. Granado-Criado, “Recommender system implementations for embedded collaborative filtering applications,” Microprocess. Microsyst., vol. 73, 2020.

[29] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, “Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster,” IOP Conf. Ser. Mater. Sci. Eng., vol. 336, no. 1, 2018.

[30] P. Bholowalia and A. Kumar, “EBK-means: a clustering technique based on elbow method and K-Means in WSN,” Int. J. Comput. Appl., vol. 105, no. 9, pp. 17–24, 2014.

[31] S. Mirjalili and A. Lewis, “The Whale Optimization Algorithm,” Adv. Eng. Softw., vol. 95, pp. 51–67, 2016.

Enhanced Algorithm for Reconstruction of Three-Dimensional Mesh from Medical Images using Tessellation of Recent Graphics Cards

Lamyae Miara[1], Said Benomar El Mdeghri[2], Mohammed Ouçamah Cherkaoui Malki[3]
Department of Computer Science, Faculty of Science Dhar El Mahraz
University Sidi Mohammed Ben Abdellah
Fez, Morocco

***Abstract*—The reconstruction of a 3D mesh using displacement vectors for medical images is a recent method that allows the exploitation of modern GPUs. This method demonstrated its efficiency by accelerating 3D visualization calculations and optimizing the storage process. In fact, it is divided into two main stages. The first step is the construction of a basic mesh by applying the Marching Cubes algorithm, and the second step is the extraction of the displacement vectors, which represent the details lost in the basic mesh. In fact, the Marching Cubes algorithm used to build the basic mesh suffers from some problems that we will try to overcome in this article. These problems are summarized in the ambiguity encountered during the construction of the basic mesh in some cases. Also, the resulting basic mesh must undergo modifications, in order not to have errors of form, which requires time and memory, and which gives the end a final mesh which is not optimal and even erroneous in certain situations. Our method is based on extracting the contours of the anatomy to be reconstructed from a sequence of 2D images. Each contour will be represented by a triangle. The shape of the basic mesh will then be the result of the connection of these triangles. This strategy avoids the use of the marching cubes algorithm in the reconstruction of the basic mesh in order to overcome the problems mentioned above.**

***Keywords*—3D reconstruction; medical imaging; marching cubes; displacement vectors; contour extraction; contour matching**

I. Introduction

Using medical imaging, it is now possible to visualize the patient's internal anatomy without resorting to surgery. The images obtained are 2D slice series on which it is not always easy to interpret the different problems faced by physicians. In order to overcome these problems, 3D imaging has been developed [1]. This technique will allow doctors, to visualize his patient in virtual 3D. This will bring a better representation of the internal anatomy in order to optimize the chances of a good diagnosis [2].

On the other hand, the recent technological development of devices used in the medical field has contributed to the resolution of the medical images obtained. The reconstruction of a 3D mesh of the human anatomy from these images generates a very large number of polygons. It preserves many details of the 2D image that lead to problems in the amount of information stored, and in terms of the complexity of the display of calculations for real-time visualization.

Therefore, this paper focuses on the algorithms of the 3D reconstruction of medical images using the tessellation of recent graphics cards. These algorithms, based in principle on the Marching Cubes algorithm [3] and a displacement map, have shown their efficiency by optimizing the amount of information to be stored in the 3D mesh, and by accelerating the rendering computations. These algorithms are based on a sequence of 2D medical images, which will be segmented to extract only the organ to be reconstructed. Then, the Marching Cubes algorithm is applied to these segmented images.

The size of the cubes selected in the Marching Cubes algorithm must be large to have a low-resolution basic mesh, this is the basic mesh. The algorithm developed by [4] allows us to extract the lost details from each basic mesh triangle using the information we have in the medical images. These details are stored as displacement vectors.

The Marching Cubes algorithm used to build a low-resolution basic mesh causes some ambiguity problems during the construction of the mesh in some cases, so the basic mesh has to be modified to avoid shape errors, which requires time and memory, and gives a final mesh that is not optimal and even erroneous in some situations.

The rest of this paper is organized as follows: the background and related works are reviewed in Section 2; in Section 3, our methodology is discussed; Section 4 was devoted to discussing our results, and in Section 5, we summarize our work with conclusion.

II. Related Works

Among the 3D reconstruction algorithms used in the medical field is the Marching Cubes. The latter generates a 3D polygonal object from a three-dimensional scalar field [5]. This algorithm executes this scalar field, taking eight points at a time from an imaginary cube and determines the creation polygons to represent part of the iso-surface contained in this cube. This algorithm is based on a precomputed table of 256 configurations of the number of polygons in a cube [6], treating each of the eight scalar values as a bit in an 8-bit integer. Although the MC algorithm has proven to be efficient, it suffers from several problems. Many solutions have been proposed to solve these problems.

The first problem detected in the MC algorithm is the case of ambiguities found in some configurations. In fact, the same cube can be triangulated in several ways. The author in [7] have identified the problem of face ambiguity, which occurs when two diagonally opposite vertices are labeled positive and two negatives are labeled negative. This ambiguity can lead to holes in the topology. The authors in [8] and [9] independently show another type of ambiguity that occurs inside a cube. These ambiguities can often be resolved by adding additional test points in each cube. Several methods are proposed to solve this problem, either for ambiguous cases on faces [7] or inside cubes [10] and [11].

The Marching Cubes algorithm divides the space into a regular cubic grid. The resolution of the resulting polygonal surface depends directly on the size of the grid. Increasing the number of cubes in the grid can increase the resolution of the polygonal area, but the number of resulting triangles can be large even if the original area is quite simple. To reduce the number of triangles, several methods have been developed to apply the Marching Cubes algorithm of the adaptive grid. A method that introduces the concept of multi-resolution grid generation [12], an algorithm that adapts the size of the triangles to the shape of the surface, has been proposed by [9]. The author in [13] proposed another method that keeps the same grid size and increases the resolution of the moving surface from the grid peaks to the surface of the cube, thus increasing the number of cubes containing the surface.

In all the proposed solutions to improve the Marching Cubes algorithm, we always have the problem of the huge mesh obtained at the end of this treatment, which directly influences the fluidity of the visualization of this mesh in real time. Among the algorithms proposed to solve this problem, we have the algorithm proposed by [14] which is based on the Marching Cubes and on the tessellation of recent graphics cards. This algorithm has proven its interest in optimizing the amount of information to store for 3D mesh and to speed up display calculations using recent GPUs.

This algorithm uses a sequence of medical images. These images are segmented to extract only the organ to be reconstructed. Then we apply, on these segmented images, the Marching Cubes algorithm [15].

The size of the cubes chosen in the Marching Cubes algorithm or its extensions must be important to have a low-resolution mesh, it is the basic mesh.

The algorithm developed by [14] extracts the lost details of each polygon from the basic mesh using the information that we have in medical images. These details are saved as displacement vectors.

The method used in the work of [14] causes some problems in the precision of the final high-resolution mesh.

Indeed, the use of Marching Cubes algorithm in the reconstruction of basic mesh always poses some problems:

- The first problem posed by this algorithm is the cases of ambiguities found in certain configurations. For the same cube we can have several ways of triangulating (Fig. 1).

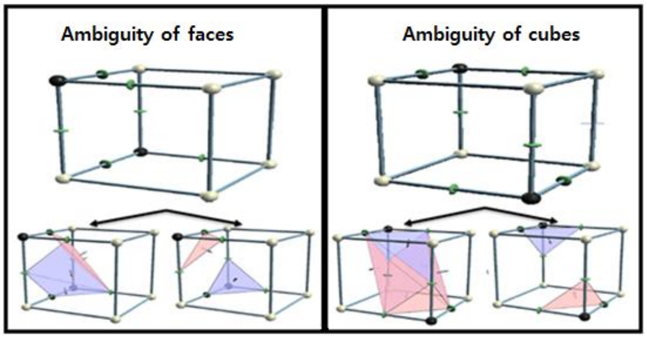


Fig. 1. Ambiguity of Faces and Cubes. Several Triangulations possible for the same case.

These ambiguities can cause inconsistencies, such as holes in the basic mesh (Fig. 2).

In Fig. 2, we represent two neighboring cells. The left cell is triangulated, the voxels having the value 1 in a separate way (case 1), the neighboring cell on the right is triangulated, the voxels have the value 1 in a connected way (case 2). Both triangulations are valid, but they generate a triangular interface incompatible with a hole. The dotted lines indicate respectively the edges of the triangle in the other cell.

- The second problem detected is the need to modify the mesh obtained by Marching Cubes in order to be used in Displacement Mapping. This treatment consists in removing some vertex from the mesh and merging triangles to obtain quadrilaterals. The application of this treatment on certain areas, indicated as complicated, leads to a deformation of the basic mesh (Fig. 3).

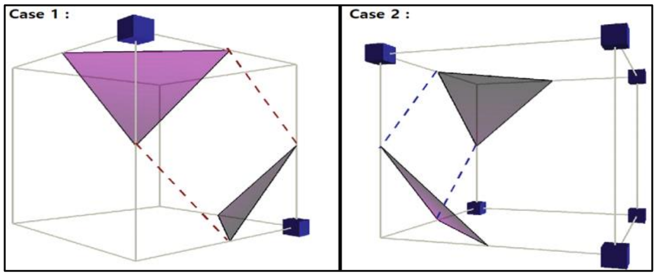


Fig. 2. The Possible Triangulations of Marching Cubes for Two Neighboring Cells.

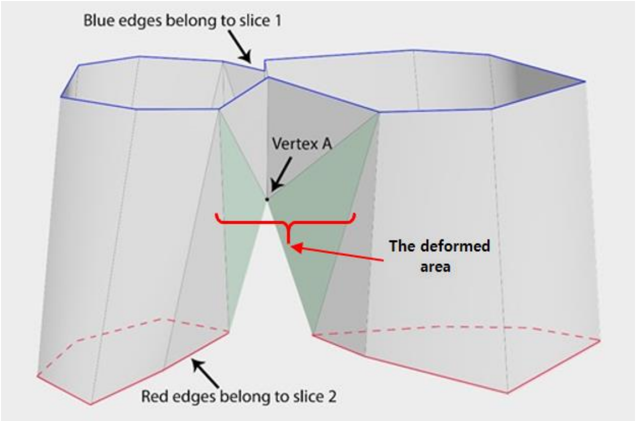


Fig. 3. Complicated Area in a basic Mesh.

In this figure, we have a mesh obtained after the application of Marching Cubes on two slices. Vertex A is located in a complicated area, the removal of this vertex leads to the deformation of this mesh. The solution adapted by the original article is not to change its position, which causes the existence of triangles (the two green triangles) in which we cannot add the details lost with displacement mapping.

In this paper, we proposed a new algorithm to reconstruct the low-resolution mesh. This technique is essentially based on an algorithm that extracts the contours and their correspondences to form this mesh. Our method overcomes the problems associated with the use of Marching Cubes.

III. Methodology

This section describes the method developed in this article to reconstruct a basic mesh without using the Marching Cubes algorithm. This method consists in extracting, from a sequence of 2D medical images, the contours of the anatomy to be reconstructed, and then to represent them by triangles, and in the final step we connect these triangles to form the low-resolution mesh.

In the first part, we determine the input of our algorithm and explain the basic idea. The other two parts describe in more detail the main steps of the algorithm: contour extraction and linking the triangles (the contour points). A last part is here to show how to build the high-resolution mesh.

A. *Overview of the Algorithm*

To overcome the above-mentioned problems, we have chosen a new strategy to build the basic mesh. This strategy is based on the extraction of the contours of the anatomy to be reconstructed from a sequence of 2D medical images. This contour will be divided into three equal parts in order to deduce the coordinates of the three points that will represent this contour by a triangle. The basic mesh will be formed by the correspondence of the triangles of the image N with the image N+1. The positions of the points of the sides that exist between two 2D images are estimated by applying the principle of interpolation, because we have no real information between 2D images (Fig. 4). The result of this processing represents the basic shape of the element to be reconstructed.

The next step is to extract the lost details from each triangle of the basic mesh using the information we have in these 2D images. These details are recorded as displacement vectors. And using this method we can also automatically generate the displacement map for our basic mesh (Fig. 5).

The inputs of our basic mesh construction method are the same inputs as the original algorithm. They are the 2D medical images undergoing a segmentation to extract only the part that we are concerned (Fig. 6).

B. *Contour Extraction*

Based on the segmented slices, we must extract the contours of the object to be reconstructed.

To detect these contours, several methods are possible, grouped into several classes, those based on non-linear filtering such as the median filter -and more recently- [16], those using high-pass filtering, such as the Prewitt, Sobel and Canny detectors [17], those of multi-scale analysis developed with wavelet theory [18] and [19], and those based on the rare approximation by redundant dictionary [20].

The previously mentioned methods of contouring consist of defining abrupt changes in pixel intensity. However, our issue recommends that we need to get a chained list of contour points, respecting a fixed order. So, the strategy we used is to extract one point from the considered contour, and then we must extract the other points in an orderly way by following the path of these contours.

In the next step we will represent each contour by a triangle. We then calculate the center of gravity of each contour. The center of gravity will be the average of the points that make up the contour (Fig. 7). The horizontal projection of this point on the contour gives at least two points. The point with the maximum abscissa value will be the first vertex of the triangle, i.e. point A. The two vertices of the triangle, the remaining point B and C, must be separated by the same number of contour points.

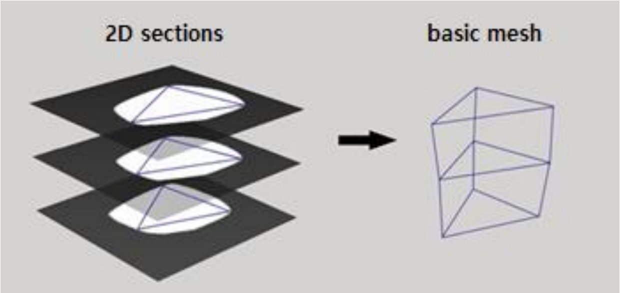


Fig. 4. The Result of the First Stage of the Reconstruction of a basic 3D Mesh from Medical Images.

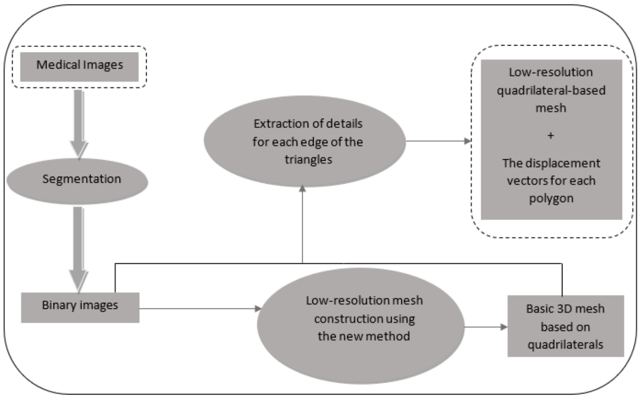

Medical Images
Segmentation
Binary images
Extraction of details for each edge of the triangles
Low-resolution quadrilateral-based mesh
+
The displacement vectors for each polygon
Low-resolution mesh construction using the new method
Basic 3D mesh based on quadrilaterals

Fig. 5. Diagram Representing the different Steps of our Algorithm.

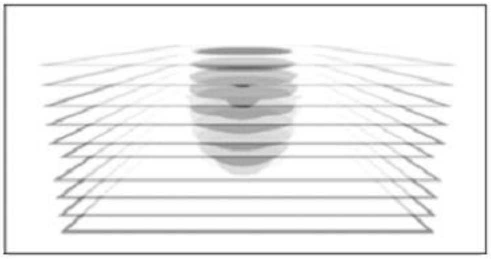


Fig. 6. The Inputs to our Algorithm, the Medical Images [14].

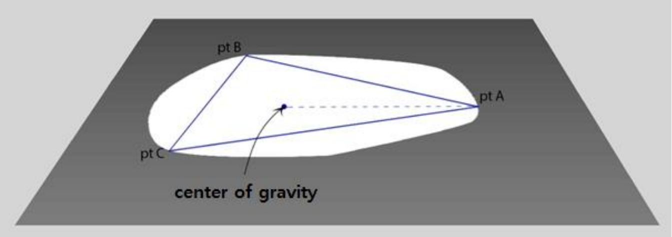


Fig. 7. Calculation of the Center of Gravity and Three Points that represent an Outline of a 2D Image.

Each contour of slice N will be represented by three points forming a triangle. These will be the basis of the low-resolution mesh of the anatomy to be reconstructed.

C. Construction of the basic Mesh

1) Contour matching: In our method, the construction of a basic mesh consists in making the correspondence between the contours of each slice n with those of slice n+1[21].

Drawing faces from the contours isn't an easy thing, since it depends on solving three problems [22]:

- Matching problem: How to match the contours in the N slice with a contour of the n+1 slice?
- Tiling problem: How to connect the points of the contour Cn in the n slice with the points of the contour Cn+1 in the n+1 slice?
- Connection problem: How to divide the contour Cn in slice n that corresponds to the contours Cn+1a and Cn+1b in slice n+1?

The representations shown below, highlights feasible possibilities to solve the matching problem in the case where the number of contours in adjacent slices is not the same (Fig. 8) (contours can be split or merged). Although the matching problem depends on the connection issue, the assumption that it can occur in isolation is not ruled out; especially if the contour curves change strongly between adjacent slices.

To determine the relationship between the contours of two consecutive slices, we used a correspondence factor. To calculate this factor, each contour is represented by a binary matrix of the same size as the slice. The pixels that are inside the contour will be represented by the value 1 and the others by the value 0 (Fig. 9).

Thus, to determine the value of the correspondence factor between 2 contours C1 and C2 of two consecutive slices (Fig. 10), we apply the logical AND operator on the matrices of the two contours.

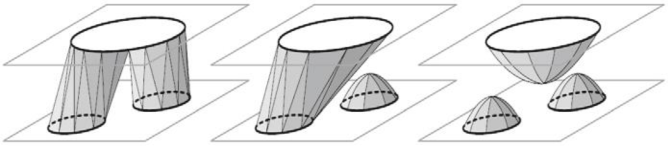
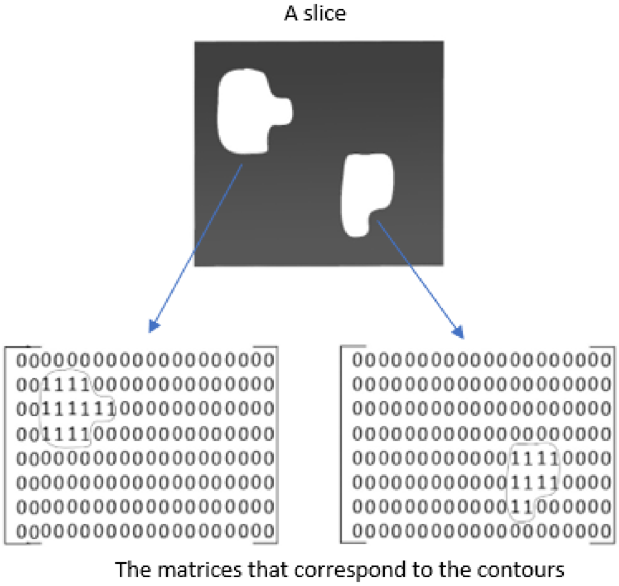
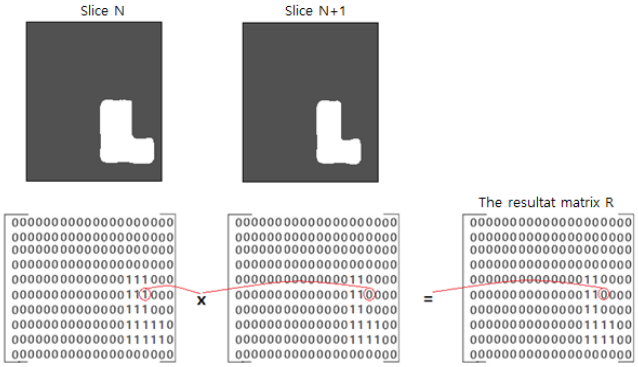


Fig. 8. Three Possible Solutions to the Problem of Matching in the Event of a Change of Topology.



```
00000000000000000000    00000000000000000000
00111100000000000000    00000000000000000000
00111111000000000000    00000000000000000000
00111100000000000000    00000000000000000000
00000000000000000000    00000000000011110000
00000000000000000000    00000000000011110000
00000000000000000000    00000000000011000000
00000000000000000000    00000000000000000000
```

Fig. 9. Representation of Matrices for the Contours of a Slice.



```
00000000000000000000     00000000000000000000     00000000000000000000
00000000000000000000     00000000000000000000     00000000000000000000
00000000000000000000     00000000000000000000     00000000000000000000
00000000000000000000     00000000000000000000     00000000000000000000
00000000000000111000     00000000000000110000     00000000000000110000
00000000000000111000  x  00000000000000110000  =  00000000000000110000
00000000000000111000     00000000000000110000     00000000000000110000
00000000000000111110     00000000000000111100     00000000000000111100
00000000000000111110     00000000000000111100     00000000000000111100
00000000000000000000     00000000000000000000     00000000000000000000
```

Fig. 10. Calculation of the Result of Two Binary Matrices.

The correspondence factor of a contour is then the sum of the elements of the resulting matrix, divided by the sum of the elements of the matrix of the same contour.

We deduce that the correspondence factor of the contour C1 in relation to C2 is:

$$Fc(C1/C2) = \frac{\sum elt\ R}{\sum elt\ C1} = \frac{4}{6}$$

and the correspondence factor of the contour C2 in relation to C1 is:

$$Fc(C2/C1) = \frac{\sum elt\ R}{\sum elt\ C2} = \frac{4}{4}$$

So, to determine the relationship between the contours of slice N and the contours of slice N+1, using the correspondence factor, there are two possible cases:

- Contour C1 of slice N is connected to only one contour C2 of slice N+1.
- Contour C1 of slice N is connected to several contours of slice N+1.

In order to verify the first case, there must be a correspondence between C1 and C2 in both directions; i.e. the contour C1 must correspond to C2, and C2 must correspond to C1 (Fig. 11). This correspondence exists when the two correspondence factors Fc (C1 / C2) and Fc (C2 / C1) are respectively greater than a given threshold. In our case, we are working on a series of slices from recent scanners, with a distance between slices in the order of millimeters. So, the correspondence factor between the linked contours is close to 100%.

In this article, we must determine a threshold for the correspondence factor (S_f) from which we can determine whether there is a correspondence between 2 contours or not. To make this choice, we have a compromise to make. On the one hand we must choose the lowest possible threshold in order to take into account all the particular cases of contour change from one slice to the other, and on the other hand, we must not have conflicts of correspondence if we accept very low correspondence factors. An example of this conflict is to have a correspondence of one contour of slice n with several contours of slice n+1. The value of the matching threshold we have chosen to meet this compromise is 50%.

In the rest of this article, we assume that the C1 contour is related to a C2 contour if the two correspondence factors are greater than 50%. This assumption will not be true if there is a large distance between the slices. The thing that is not valid in our situation.

And for the second case, we have one contour of slice N which corresponds to two or more contours (C2, C3… CK) of slice N+1. In this case, the correspondence factor is checked only in one direction (Fig. 12).

In this figure, the contours of slice N+1 correspond to the same contour of slice N. The correspondence factors Fc(C2/C1) and Fc(C3/C1) are greater than the determined threshold. However, the correspondence factors Fc(C1/C2) and Fc(C1/C3) may be necessarily lower than the determined threshold.

In the special case, where there are two contours in the same slice N which correspond to each other, and which correspond to the same contour in slice N+1 (Fig. 13), one contour must be removed from slice N to take off this correspondence. In fact, the stronger match is the one that will remain. Also, since Fc(C2/C3) + Fc(C1/C3) is greater than Fc(C2/C1) + Fc(C1/C2), then contour C2 will be removed from this matching assumption.

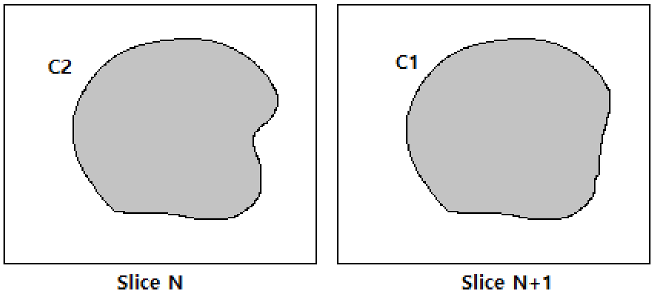


Fig. 11. Contour C1 of Slice N and Contour C2 of Slice N+1 are Linked.

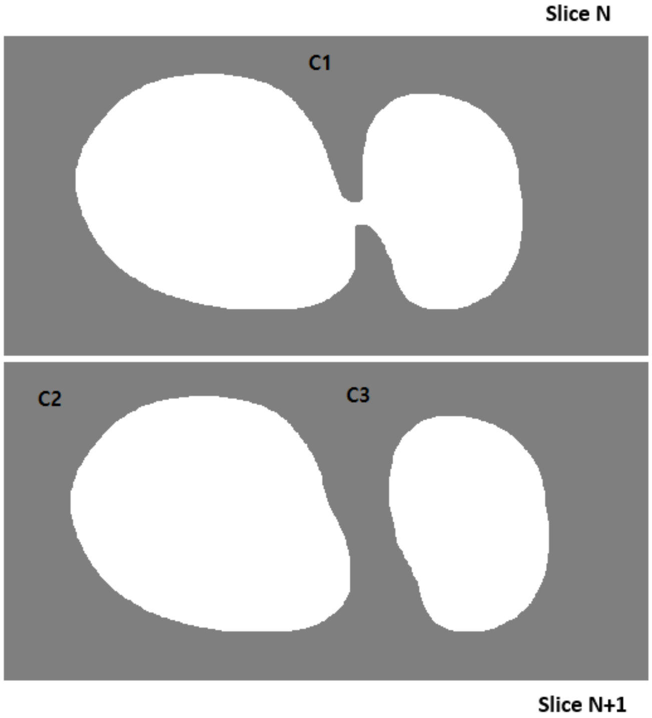


Fig. 12. The Contour C1 is Divided into Two Contours C2 and C3 in the following Slice.

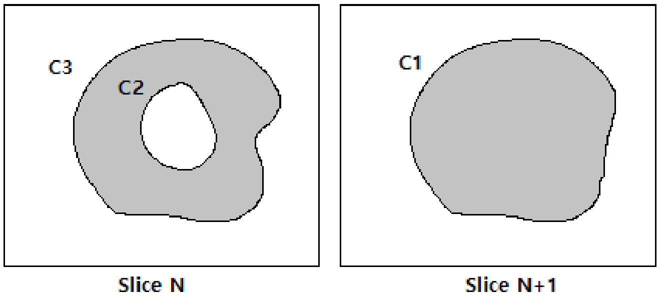


Fig. 13. Two Consecutive Slices that have Contours that Correspond to each other.

2) Construction of the low-resolution mesh: The construction of the mesh amounts to linking the corresponding contours. And to obtain a low-resolution mesh we just connect the triangles that represent the contours. In this step, there are two cases:

- A contour of slice N is connected to a single contour of slice N+1.
- One contour of slice N is connected to two or more contours of slice N+1.

In the first case, the construction of the mesh amounts to connecting the vertices (A, B, C) of the contour C1 with the vertices (A’, B’, C’) of the contour C2 respectively (Fig. 14).

In the second case, the construction of the mesh amounts to connecting the contour C1 of the slice N with the vertices ((A', B', C'), (A", B", C"),…) of contours (C2, C3,…, Ck) by taking into account the correspondence factors Fc (C1 / C2), Fc (C1 / C3),…, Fc (C1 / Ck).

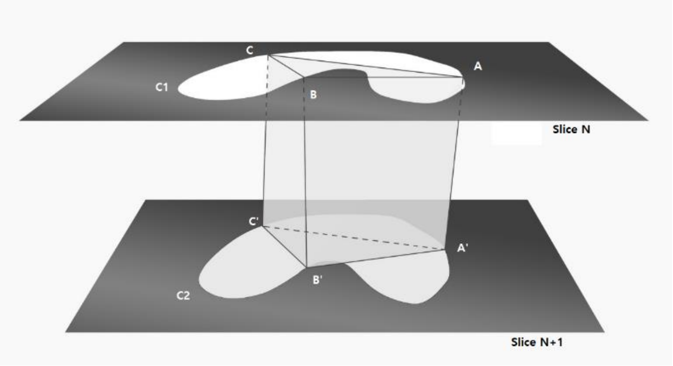


Fig. 14. Connection of Contour C1 of Slice N to Contour C2 of slice N + 1.

Contour C1 will be divided according to the number of contours that correspond to it, and according to the percentage of correspondence of contours C2, C3, ..., Ck with respect to contour C1 is calculated.

The matching factor is used to calculate the matching percentage.

We suppose that: $\sum_{n=1}^{k} Fc\left(\frac{C1}{Cn}\right) = X$

And we use the triangular relation to calculate the percentage of correspondence:

$Fc\left(\frac{C1}{Cn}\right) \rightarrow \qquad Pc\left(\frac{C1}{Cn}\right)$

X → 100%

Then, we have:

$Pc(C1/Cn) = \quad Fc(C1/Cn) \quad / \quad X$

This percentages of correspondence indicate the percentage of correspondence of the contours C2, C3, ..., Ck with respect to the contour C1. And to locate the position of a contour Ck in the contour C1, we project the center of gravity of the contour Ck on the contour C1. Next, the correspondence area is calculated using the correspondence percentage (Fig. 15). Henceforth, we can determine the three points (A, B, C) of the sub-contour which corresponds to each contour Ck.

In the figure above, the two contours C2 and C3 correspond to the contour C1. So, to calculate the correspondence surface of each contour of slice N + 1 in contour C1, we first projected the centers of gravity of the two contours on contour C1. Then, we made the intersection of the straight line (D) passing by the two projections and the contour C1. However, we calculate the correspondence percentages Pc (C1 / C2) and Pc (C1 / C3). Subsequently, we did a sweep from the end of each intersection of the contour C1 with the line (D) according to the match percentages. In Fig. 15 the sub-outline on the right (green color) represents 33.33% of the contour C1, and the sub-contour on the left (orange color) represents 66.67% of the contour C1.

The next step is to match each contour of the N+1 slice to its corresponding contour in the N slice (Fig. 16).

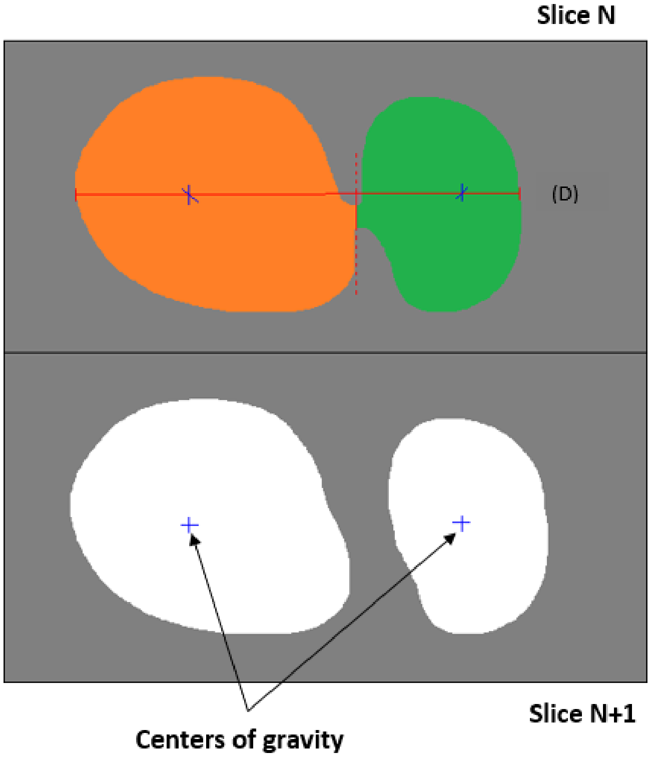


Fig. 15. The Correspondence of the Contour C1 to Two Contours C2 and C3 of the Slice N + 1.

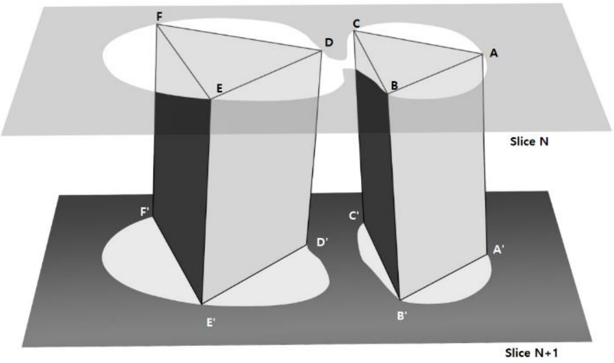


Fig. 16. Construction of the Low-Resolution Mesh from Two Contours C2 and C3 which Correspond to the Contour C1.

D. Construction of the High-Resolution Mesh

As shown in the previous diagrams, the constructed mesh is a basic mesh that does not reflect the true shape of the anatomy in question. It is necessary to add to it the details lost during the construction of the basic mesh in real-time visualization. The displacement vectors must then be extracted using the contour data and quadrilateral edges obtained after linking the triangles that represent the contours.

The detail extraction method used in the original article, is based on the discretization of the edges that are on the 2D images in N points, then to extract for each point the corresponding displacement vector.

Two problems arise when using the displacement vector extraction algorithm to obtain a high-resolution mesh; the first is that all sides of polygons are discretized to N points, with N fixed. However, the widths of the sides are not identical (Fig. 17), and therefore logically the homogeneity of the final 3D mesh will be altered with respect to the details extracted from each polygon.

It can be noted that the level of detail extracted from side A is less important than that of side B.

The second problem concerns the final shape of the anatomy to be reconstructed, which presents errors in certain situations. Drawing a perpendicular line from the points resulting from the discretization of each side sometimes gives two or more points of the intersection with the contour in 2D images. The choice of the extraction method of the original algorithm leads to erroneous results, because through this method the choice of the vector is always made between a point that belongs to the side and the closest intersection point of this side. This alters the accuracy of the result especially since the details no longer correspond to the true shape of the anatomy (Fig. 18).

We notice that the obtained shape (c), using these displacement vectors (b), does not really resemble the true shape (a).

In a more recent work by the same author [4] the extraction method has been modified to overcome the problems mentioned above.

The new strategy used is to extract the displacement vectors is based on the extraction of the contour of the object to be reconstructed from the medical images, then, using discretization of these contours we generate the displacement vectors.

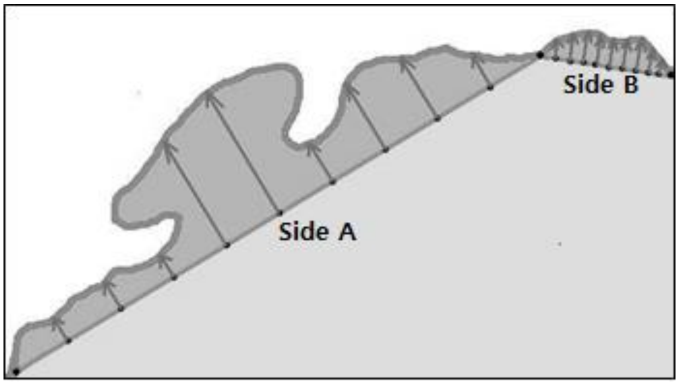


Fig. 17. Heterogeneity of Details Extracted from each Side [4].

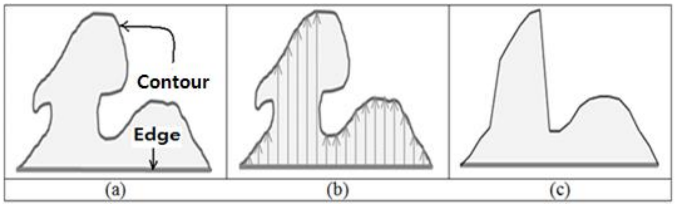


Fig. 18. Errors in Extracting the Small Reliefs. The Contour Obtained (c) by using the Displacement Vectors (b) does not Correspond Exactly to the True form (a) [4].

This new extraction method has an importance related to two main issues; the first concerns the generation of vectors considering the contour itself, this allows more precision for the construction of a real anatomical shape. The second interest is the integration of the same level of detail in all the polygons of the basic mesh, through the discretization of the contour of the anatomy with a fixed point.

Displacement vectors will be used for the automatic generation of the displacement map, which is an image allowing to determine the distance and direction used to move the points of the surface during real-time visualization.

We follow the same approach used in the original paper to generate the high-resolution mesh based on the basic mesh obtained with our new method and the displacement map (Fig. 19).

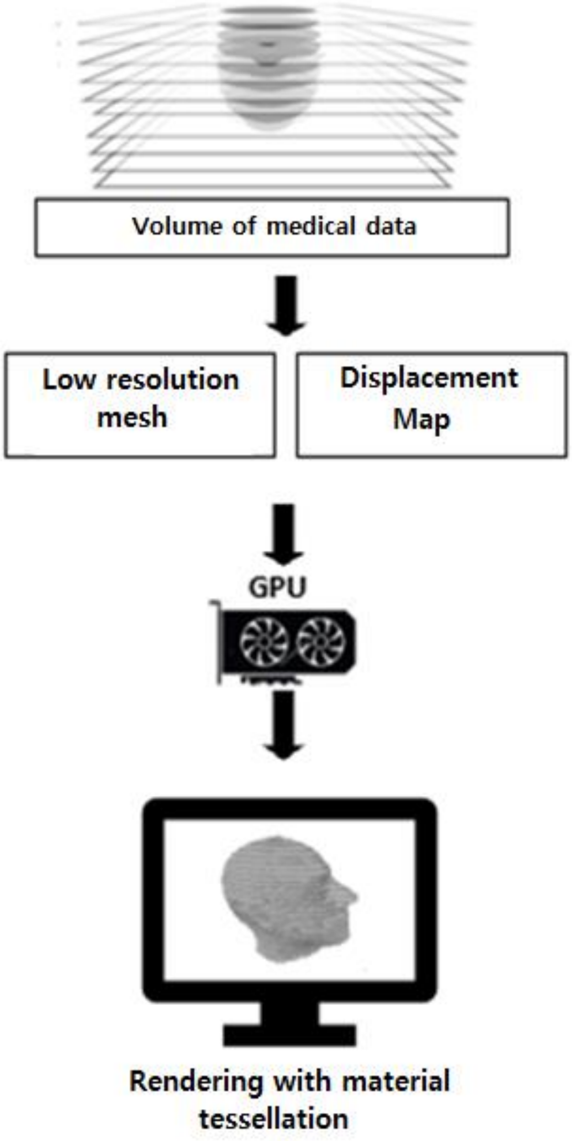


Fig. 19. Rendering with Hardware Tessellation to Generate the High-Resolution Mesh.

IV. Results and Discussion

The implementation of our new method for reconstructing a 3D mesh from medical images is based on:

- The detection and extraction of contours from a sequence of medical images;
- The construction of a low-resolution 3D mesh from the contours;
- The construction of the high-resolution mesh from the displacement vectors and the basic mesh.

The image sequence used in this paper is segmented to extract only the anatomy to be reconstructed.

Initially, we tested our new 3D mesh reconstruction method in terms of the quality of the obtained 3D mesh.

The basic mesh construction method used by the original algorithm is based on the use of the Marching Cubes algorithm, while determining the size of the cubes. This method causes cases of ambiguity in some configurations; i.e. for the same cube, there can be several ways of triangulation. Also, with this method, an additional step is mandatory to render the mesh obtained from the Marching Cubes algorithm in the form of quadrilaterals, in order to apply displacement mapping. And in some situations, complicated areas are confronted during this step, which cause deformations in the reconstruction of the high-resolution mesh.

In Fig. 20, point A is located in a complicated area. This point will lead to deformation when applying displacement vectors, since there is no information at this point.

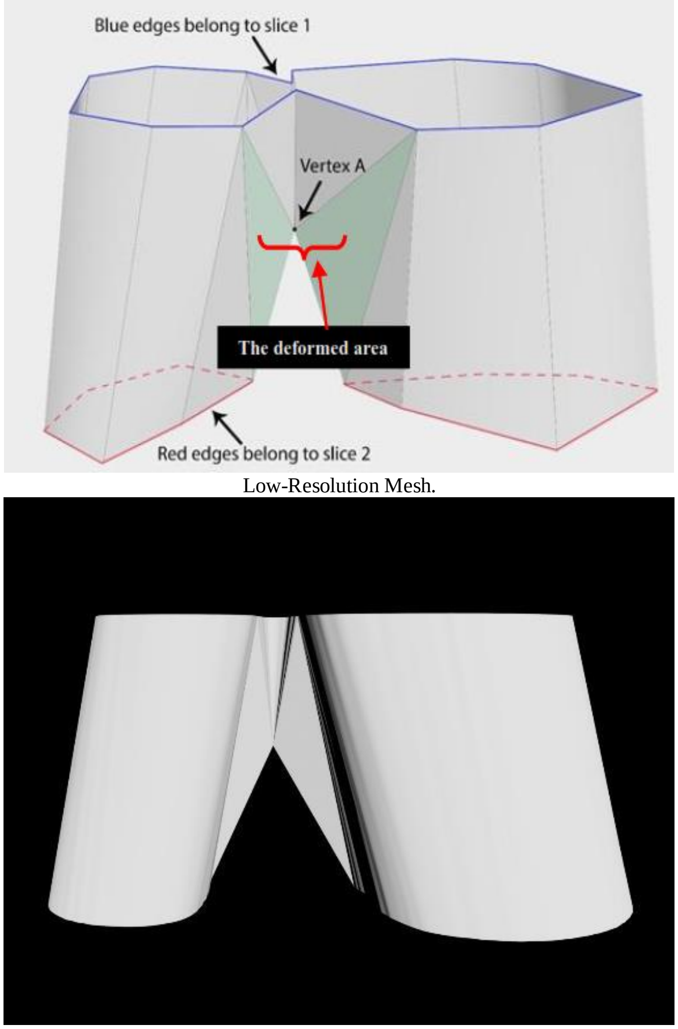
In our new method, the extraction of the contours and the use of the correspondence factor allows the construction of a 3D mesh without ambiguity, and directly in the form of quadrilaterals.

In fact, after extracting the contours in the 2D images, the correspondence factor for the contours of each image N is calculated with the image N+1. This factor makes it possible to indicate the contours that match each other unambiguously. Also, the construction of the basic 3D mesh is done by representing each contour by a triangle. Then, the triangles of the contours that correspond to each other are connected. This connection results in a mesh in the form of quadrilaterals (Fig. 21).

We also note that our new method incorporates an improvement made to this article in terms of displacement vectors [4] (Fig. 22).

Note that the mesh obtained with high resolution corresponds to the real anatomical shape.

Subsequently, we tested our new method at the storage level. The algorithm used in the original method allows us to generate a basic mesh with many vertices. But with our 3D mesh construction method which is based on contour extraction and the representation of each contour by a triangle, we notice that the amount of storage is less important (Fig. 23 and Fig. 24).



Low-Resolution Mesh.

High-Resolution Mesh.

Fig. 20. The Modification Step Leads to Deformations in Complicated Areas.

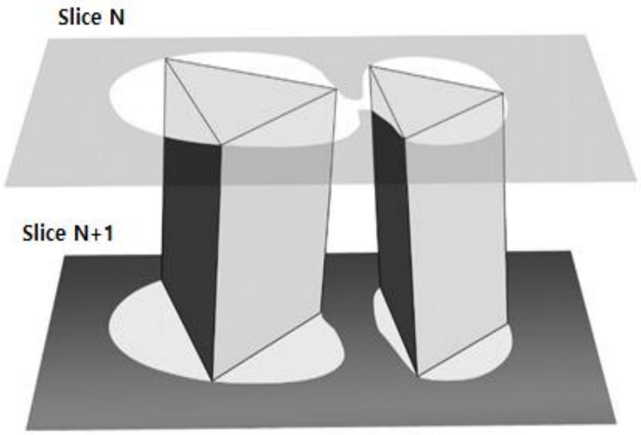


Fig. 21. A basic Mesh in the form of Quadrilaterals with our New Method.

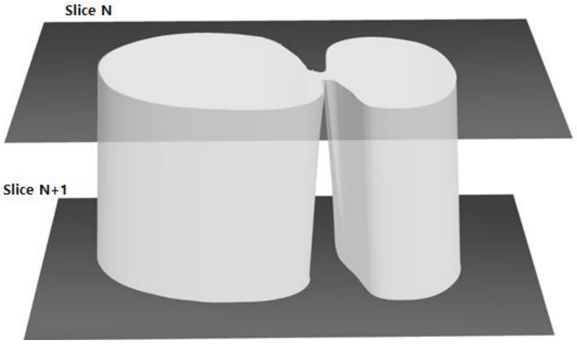


Fig. 22. The High-Resolution 3D Mesh Obtained after the Application of the Displacement Vectors by Integrating the Improvement.

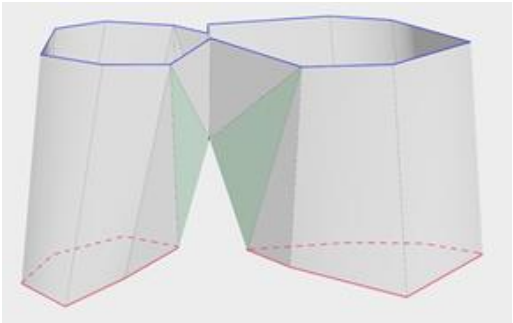


Fig. 23. Basic Mesh with the Original Method.

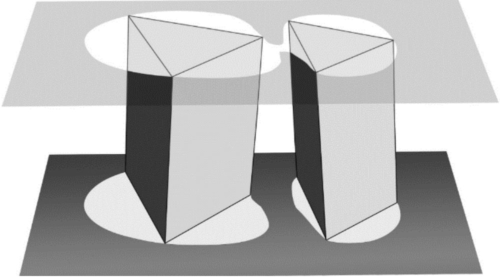


Fig. 24. Basic Mesh with our New Method.

If we compare the used memory space of the two 3D meshes between two 2D images (Table I), we find:

We worked on a medical volume consisting of 30 slices. With the original method, we obtained a basic mesh size of 40 501 bytes, and with our method, we obtained a basic mesh size of 37 675 bytes.

We can see from these calculations that our new method is optimized in terms of storage.

TABLE I. CALCULATION AND COMPARISON OF THE SIZE OF MESH IN THE ORIGINAL METHOD AND THE NEW METHOD

	Basic mesh between two 2D images	**Displacement vectors**	**Total:**
The original method	27 vertex *3 float(x,y,z) *4 octets	500 vectors *2 char(x,y) *1 octet	=1324 bytes
The new method	12 vertex *3 float(x,y,z) *4 octets	500 vectors *2 char(x,y) *1 octet	=1144 bytes

V. CONCLUSION

In this paper, we proposed an improvement of the algorithm that allows the reconstruction of the 3D mesh for a 2D medical image sequence, using low-resolution mesh and displacement vectors. The use of displacement vectors to add small reliefs on a basic mesh has proven to reduce the amount of information stored in the final 3D mesh. Also, the automatic generation of the displacement map speeds up rendering calculations using the GPU.

In this work, we modified the method used to build the basic mesh. This modification aims to eliminate the cases of ambiguity, particularly in certain types of objects which cannot be treated easily with the original method which uses the algorithm of Marching Cubes. Thus, it also makes it possible to obtain a low-resolution mesh directly usable during rendering without going through a modification step.

In our new method, we extracted the contour of the anatomy to be reconstructed for the 2D images, then we built a 3D mesh in minimal time, and optimized in terms of memory. Then we generated displacement vectors by discretizing the contour according to the level of detail we want to see.

Our method has proven itself on two main points:

- The reconstruction of the low-resolution 3D mesh using triangles that represent the contours provides a basic mesh without deformation problems that can be used in rendering.
- The edge detection method allows to overcome the ambiguity problems from which the Marching Cubes algorithm suffers.

REFERENCES

[1] J. Tan, J. Chen, Y. Wang, L. Li, and Y. Bao, “Design of 3D Visualization System Based on VTK Utilizing Marching Cubes and Ray Casting Algorithm,” 2016.

[2] Bücking, T. M., Hill, E. R., Robertson, J. L., Maneas, E., Plumb, A. A., & Nikitichev, D. I.,” From medical imaging data to 3D printed anatomical models. Plos One,” 12(5), 2017.

[3] W. E. Lorensen, "History of the Marching Cubes Algorithm," IEEE Computer Graphics and Applications, Vol. 40, No. 2, pp. 8–15, 2020.

[4] B.E. Said, M.O.C. Malki. and K. Abdelhak, “Improvement of the generation of displacement vectors in the reconstruction of a 3D mesh for medical images,” Int. J. Medical Engineering and Informatics, Vol. 9, No. 1, pp.20–21, 2017.

[5] S .Mady and M. El Seoud, “An Overview of Volume Rendering Techniques for Medical Imaging, ” International Journal of Online and Biomedical Engineering, Vol. 16, No. 6, pp 95–106, 2020.

[6] P. Visutsak, “Marching Cubes and Histogram Pyramids for 3D Medical Visualization,” Journal of Imaging, 6(9), 88, 2020.

[7] G. M. Nielson and B. Hamann, “The asymptotic decider: resolving the ambiguity in Marching Cubes,” in Proc. of IEEE Visualization, pp.83–91, 1991.

[8] B.K. Natarajan, “On generating topologically consistent isosurfaces from uniform samples,” The Visual Computer, Vol. 11, No. 1, pp.52–62, 1994.

[9] E. Chernyaev, “Marching Cubes 33: Construction of Topologically Correct Isosurfaces,” (No. CERN-CN-95-17), 1995.

[10] A. Lopes and K. Brodlie, “Improving the robustness and accuracy of the marching cubes algorithm for isosurfacing,” IEEE Transactions on Visualization & Computer Graphics, Vol. 9, No. 1, pp.16–29, 2003.

[11] G. M. Nielson, “On marching cubes,” IEEE Transactions on Visualization & Computer Graphics, Vol. 9, No. 3, pp.283–297, 2003.

[12] R. Shu, Z. Chen and M.S. Kankanhalli, “Adaptive marching cubes,” The Visual Computer, Vol. 11, No. 4, pp.202–217, 1995.

[13] J. Congote, A. Moreno, I. Barandiaran, J. Barandiaran and O. Ruiz, “Extending marching cubes with adaptative methods to obtain more accurate iso-surfaces,” Computer Vision, Imaging and Computer Graphics. Theory and Applications Communications in Computer and Information Science, Vol. 68, pp.35–44, Springer, Berlin, Heidelberg, 2010.

[14] B.E. Said, M.O.C. Malki, K. Abdelhak, and E. Abdelali, “Reconstruction of a 3D mesh with displacement vectors for medical images,” Int. J. Medical Engineering and Informatics, Vol. 7, No. 3, pp.209–221, 2015.

[15] E. Lorensenw and H. E. Cline, “Marching cubes: a high-resolution 3D surface construction algorithm,” Proc. of ACM SIGGRAPH, pp.163–169, 1987.

[16] O. Laligant, and F. Truchetet, “A nonlinear derivative scheme applied to edge detection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, February, Vol. 32, No. 2, pp.242–257, 2010.

[17] J. Canny, “A computational approach to edge detection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI, November, Vol. 8, No. 6, pp.679–698, 1986.

[18] S. Yi, D. Labate, D. Easley, and H. Krim, “A shearlet approach to edge analysis and detection,” IEEE Transactions on Image Processing, May, Vol. 18, No. 5, pp.929–941, 2009.

[19] L. Zhang and P. Bao, “Edge detection by scale multiplication in wavelet domain,” Pattern Recognition Letters, Vol. 23, No. 14, pp.1771–1784, 2002.

[20] J. Mairal, M. Leordeanu, F. Bach, M. Hebert and J. Ponce, “Discriminative sparse image models for class-specific edge detection and image interpretation,” in Forsyth, D., Torr, P. and Zisserman, A. (Eds.): Computer Vision: ECCV 2008, Lecture Notes in Computer Science, Vol. 5304, pp.43–56, Springer, Berlin, Heidelberg, 2008.

[21] R. Mukundan, “Reconstruction of High-Resolution 3D Meshes of Lung Geometry from HRCT Contours,” IEEE International Symposium on Multimedia, pp 247–252, 2016.

[22] D. Meyers, S. Skinner and A. K. Sloan, “Surfaces from contours,” ACMTransactions on Graphics, 11(3), p. 228–258, 1992.

Determinants towards a Better Acceptance Model of IoT in KSA and Eradication of Distrust in Omnipresent Environments

Abdulaziz A. Albesher[1]
College of Computing and Informatics
Saudi Electronic University, Riyadh, Saudi Arabia

Adeeb Alhomoud[2]
College of Science and Theoretical Studies
Saudi Electronic University, Riyadh, Saudi Arabia

***Abstract*—This paper highlights several of the key determinants that play a vital role in the acceptance of Internet of Things (IoT) technologies in the Kingdom of Saudi Arabia (KSA). Based on the governmental focus towards technology and the response of the citizens towards embracing new technologies, several determining factors are presented. Certain essential application areas of IoT are analyzed including the local industry, agriculture and livestock, health, education, smart metropolitans and smart government. In addition, we also explore acceptance at the personal level, such as home and privacy of individuals, security, and personal management with IoT wearables. Towards the end of this paper, some challenges of the IoT acceptance are presented along with the analysis of key enablers. All the rationalizations lead to the conclusion that IoT acceptance is inevitable based on the number of associated benefits which will enhance once the posed challenges are addressed.**

Keywords*—*Internet of Things (IoT); security; health; IoT acceptance; smart cities

I. Introduction

The Internet of Things (IoT) is one of the most accelerated emerging technologies that is perceived to be widely spread in the coming decade [1]. It comprises of an expanding network of smart devices that are capable of communicating over the internet and use various network services to interact with other internet-enabled devices. IoT is being applied in various domains including health and medical care [2], smart organizations [3], smart homes, smart cities, agriculture [4], e-commerce and personal management, to name a few. In addition, it is expected to be a catalytic agent for a variety of other advancements such as upgrading of the existing broadband networks [5], advancement in sensor technologies, longer battery life, emergence of situation-aware [6] and business-aware organizational networks [7], and many more like the same. It is estimated that the number of IoT-enabled devices connected to the internet would reach over 43 billion by 2023 [8]. There are a number of benefits with the emergence and blending of IoT technologies in our daily lives. In the corporate world, IoT is an enabler for business digitalization strategies [9]. Data gathered from the IoT devices is highly useful for analyzing consumer behavior, choices, consumption and attitudes. This can increase a company's competitive advantage by transforming its services and products into higher amplifications that are never seen before. Likewise, IoT is transforming the agricultural economies of the world by employing online monitoring, efficiency, cleaner processes, reduced resources and greenhouse automation for better crop yield [10]. IoT is also adopted for automating several of the civic amenities in smart cities [11]. In smart homes and buildings, IoT contributes to facilities management, energy management, occupants and resources tracking, and comfort enhancement. On the frontiers of health and automatic disease classification [12], IoT has bolstered through the concepts of ambient-assisted living, wearable devices, internet of mobile things and similar health information systems [13][14]. In the education sector, IoT has influenced with sensors, intra-communication among wearable technologies, augmented reality and cloud computing for assisted and remote learning [15].

Along with many potential benefits, there are several risks and elements of distrust associated with IoT, which hinders the acceptance of IoT at the personal and national levels. The first and the foremost risk element associated with IoT is the security of the data. As presented in a recent survey [16], a large number of IoT devices are still prone to hacking. As IoT technologies provide communication and access on-the-go, privacy of the individuals is also at higher stake. This leads to a distrust in the technology and could also pave path to the rights infringement [17]. In addition, some IoT environments also pose threats to health via excessive radiation [18].

Based on these potential benefits and associated risks perceived by the masses, countries around the world face reluctance and insecurity in the acceptance of IoT technologies. There have been several models discussed in literature for IoT acceptance. Some prominent models among these include Technology Acceptance Model (TAM), Theory of Reasoned Action (TRA) [19], and Structural Equation Model (SEM). In order to appropriately utilize any of these models, it is equivalently important to select the right set of attributes which are deterministic.

In this work, we exploit several of the determinants that play a critical role in IoT adoption and in structuring of a model that is highly judicious for IoT acceptance in the Kingdom of Saudi Arabia (KSA). Based on extensive literature survey, we investigate several key application areas of IoT, which include acceptance at the national level, such as in the industry, health, education, civic amenities and agriculture. In addition, we explore acceptance at the personal level, such as home and privacy of individuals, security, and personal management with IoT wearables. Towards the end of this study, we highlight some of the enablers of IoT acceptance and various challenges faced that can hinder its acceptability.

The rest of the paper is organized as follows. Section 2 discusses the factors affecting IoT acceptance in the economic and societal sectors in Saudi Arabia. It also presents and analyzes some of the determinants that contribute to the acceptance of IoT at the personal levels. Section 3 discusses IoT acceptance enablers and the challenges that hinder the potential growth in KSA with the implementation of IoT. Conclusions are presented in Section 5, followed by the references.

II. IoT Acceptance at the National Level

There are a number of application areas of IoT at the national level in KSA. We segregate them in to five major areas including industry, agriculture, health, education and civic amenities. The following subsections highlights the key determinants of IoT acceptance in these areas.

A. IoT in Industry

The industrial sector of any country is one of the foundation pillars that bear the weight of major economic growth. There are several benefits pertaining to IoT integration that motivates its acceptance in the industrial and corporate organizational setups in KSA. Owing to the fact that IoT technologies dispense a two-fold communication possibility by providing people-to-things as well as things-to-things communication, its adoption in the industry is highly desired. From the point of view of industrial ecosystem [20], IoT could prove to be a central part that is enticing for the ecosystem actors to engage with. These actors, or determinants, may include:

- Supply Network: The products and services offered by a specific industry. It could be in the form of knowledge, services or goods.
- Influencers: Experts, standards alliances, industry associations, social media bloggers, journalists, worker unions etc. who can influence IoT acceptance.
- Complementors: The additional services or products offered along with the supply network.
- Intermediaries: Retailers, distributors and similar entities that facilitate the business and acts as a bridge between the industry and the end users.

In addition to these, IoT technologies are particularly very beneficial in implementing industrial safety. Industrial accidents can lead to crucial loss of human life and expensive equipment. There are a variety of sensors and actuators that are eminently useful for maintaining safety and high standards in industrial production environments. These include sensors for measuring temperature, touch, pressure, humidity, smoke, proximity, gas leakage, to name a few. Likewise, distances can be detected and accidents can be prevented using various ultrasonic and infrared sensors.

Through Wireless Body Area Networks (WBANs) [21], employees' health can be continuously monitored. Any ailment, lack of conscience or abrupt changes in vital measurements such as blood pressure, heart rate or glucose levels can be communicated to a centralized system for responding and taking preventive measures.

Other determinants of IoT acceptance in the industry pose various kinds of challenges such as devices challenges, network challenges and data challenges [22]. Devices in a network may have different capabilities and communicate over different protocols. These capabilities are both related to the hardware and the software. Non-standardized naming of devices in a network is another challenge. The integration of cloud computing with IoT expands the operating scale of industrial applications. Data in the industrial application of IoT can be categorized as (1) Data by the things, such as the values of temperature, distances, pressure, etc.; and (2) Data about the things, such as IDs, names, addresses, types etc. One of the examples of such application is in churn prediction [23]. The huge amounts of data generated from the IoT-enabled devices helps in better analysis to predict users likely to churn. Hence, the industrial acceptance of IoT is largely dependent on the afore-mentioned determinants that play a vital role.

B. IoT in Agriculture and Livestock

Like all the major countries of the world, KSA also produces most of the food indigenously, besides lack of water and extremely hot and dry climate. In circumstances such as these, IoT is a promising technology for the growth of agriculture in terms of yield, crop management and disease eradication. Data is gathered and shared over the network using a number of different sensors such as soil sensors, leaf sensors, stem sensors, roots sensors, temperature, humidity and fruit size sensors [24]. Besides these sensors, precision based devices such as drones and robots are also utilized for smart and precision agriculture.

There are a number of enablers that determine the use and acceptance of IoT technologies in agriculture and livestock. These include soil and water quality monitoring, farm monitoring, irrigation and nutrition management, disease and pest monitoring, crop health monitoring, cattle movement and management, controlled use of fertilizers, farm assets tracking, and intelligent greenhouses [10]. Fig. 1 shows a typical scenario of a farm area network based on IoT.

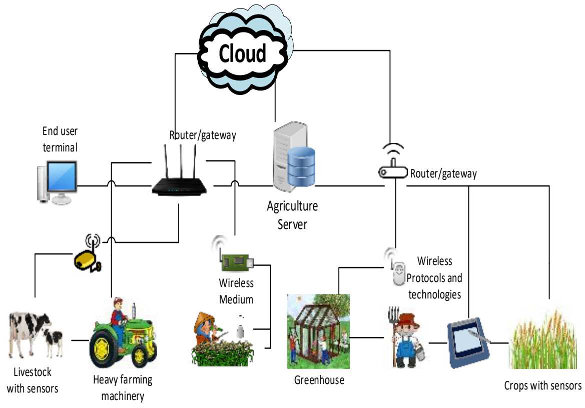


Fig. 1. A Typical Farm Area Network based on IoT Technology.

Referring to the scenario of KSA, there a number of application areas that can highly benefit from IoT technologies, thus indicating the likelihood of its immediate acceptance. These include:

- Irrigation.
- Fertilizer Management.
- Soil sampling and mapping.
- Pest management and disease control.
- Yield monitoring and harvest forecasting.
- Farm management.
- Cattle produce management.

Apart from huge potential of IoT in agriculture sector, there are a number of challenges. These include the network and hardware challenges, sensors and battery life, architectural platform differences, interference, reliability and scalability of IoT based networks. These challenges are also key determinants in IoT acceptance in the agricultural sector in KSA.

C. IoT in Education

During the recent outbreak of COVID-19 pandemic, almost all of the educational institutes around the globe either suspended their educational services or shifted to online education. It is evident that IoT technologies possess a great potential to be used in a variety of ways in order to help both the teachers and learners at the same time. One of the prime determinants in its acceptance is the enhanced learning and teaching experience derived from the analysis of data collected from various devices and sensors. Among many possibilities of enhancement of the education sector with IoT in KSA, some of the prime benefits are reducing the dropout rate in the examination through superior learning experience, achieving the learning objectives with ease, and improving the overall operations of the institutions.

The three essential skills devised for the students in the 21st century include learning and innovation skills, life and career related skills, and information technology skills [25]. IoT introduces technologies to the students, which highly inculcate in-depth understanding of the subjects of study, as well as allows their continuous physiological and behavioral monitoring. The data gathered from this can be used to design learning pedagogies that yield maximum learning capabilities with ease and less physical and mental strain.

Based on the importance of several possible factors that contribute to the overall education ecosystem improvement with the use of IoT, the following are some of the most significant determinants for IoT acceptance in education:

- Trust: A trustworthy and reliable security and privacy mechanism is needed that is readily embraced and poses no infringements of learners and educators security and privacy.
- Cost: The associated cost of the IoT adoption is anticipated to be less as compared to the expected gain. A higher cost could lead to hindrance in the acceptance of IoT.
- Expected Performance Elevation: The trust level of a user that by embracing IoT, the overall performance would be elevated.
- Expected Effort (ease of use): The amount of effort and time spent in learning new IoT enabled technologies and adapting it. It is anticipated that with technologies like online learning management systems and virtual classrooms, time and effort spent in travelling and coping with the distances will be highly reduced.
- Social Influence: One of the prime factor in acceptance of IoT is how society and influencers, such as prominent personalities, government agencies and social media, react to it. The more welcoming the response would be from these actors, the higher would be the acceptance rate of IoT in KSA.
- Hedonistic Motivation: The overall experience of learning and teaching with IoT technologies is desired to increase the pleasure in learning such as to increase motivation towards learning. The higher the pleasure level, the greater are the chances of acceptance.

The above mentioned factors also possess a number of associated challenges and risks. For the acceptance of IoT, these challenges need to be addressed and risks to be mitigated [26].

D. IoT in Health

With the transformation of the wireless sensor networks into IPv6 based lower power wireless personal area networks (6LoWPAN), smallest of the devices, with inadequate processing capability and battery-life, can transmit information wirelessly over the internet. One of the appreciable implementations of these in the medical sector is in the form of smart wearables capable of measuring and communicating vital health indicators to a centralized health monitoring system, which is connected to a hospital or other emergency medical service providers. There are a considerable number of application areas of IoT in the health sector. Some of the most prominent ones include diagnostics, counselling and therapy, drugs reference, clinical communications and medical application.

With these potential benefits, there are some associated risks of IoT adoption. One of the most prominent risk factors is the security of patient's information and privacy of the data. A slightest modification of the data due to any illegitimate infringement could result in a life-and-death scenario. Other factors affecting the acceptance of the IoT devices include the precision of the sensors and electronic measurement devices, the cost of the devices, ease of use, and the battery life.

E. IoT in Smart Cities

The adaptation of IoT technologies and the availability of devices and sensors communicating among them creates an ecosystem, which leads to smart cities. Smart cities use smart things to carry out various civic functions automatically such as efficient utilization of energy, traffic control, emergency management, smart transportation and ride sharing, lighting control, health and pollution monitoring, and surveillance [27].

Like other metropolitan cities of the world, KSA has many big cities with high population. They can receive deterministic benefits from IoT technologies in a number of ways. One of the most significant data to manage could be the traffic data. With appropriate analysis of this data, municipal governments and citizens can take a number of benefits ranging from traffic management and congestion control, efficient utilization of available parking spaces, logistics management, to better utilization of urban transport systems.

Another area where the data gathered through IoT devices could play a significant role is in surveillance of the modern cities. During the past two decades, the world has witnessed major terrorist activities around the globe. Most of these activities were carried out on public places. With smart systems capable of detecting anomalies using widely spread camera networks, terrorists' activities and other crimes can be efficiently monitored.

Smart cities also provide better means of pollution control and resource management. Water and energy can be efficiently managed and consumers can be billed with different customized option through the data gathered via smart metering. This benefits both the government, in terms of efficient resource utilization, as well as the citizens, in terms of customizable utility option that minimizes bills.

Another very important factor that contributes to IoT acceptance is the transformation of ordinary infrastructures into smart homes [28], offices and buildings. Various IoT enabled sensors and devices are highly useful in improving the everyday life and freeing up residents to perform other responsibilities. These include sensors based systems for automatic security and surveillance of buildings, efficient utilization of resources such as electricity and water, and emergency management systems that are connected to the hospital, police and fire departments. Recent state of the art systems are also capable of checking and maintaining building health and environment.

Along with the aforementioned deterministic benefits of IoT acceptance, there are several challenges that are of consideration. One of the biggest challenges is the quality of data and its integrity. IoT technologies are still novice and a lot of enhancement and improvement is required to assure complete data integrity with highest quality. Another important challenge is of the management and coordination of various systems in the IoT-based smart cities. The massive and speedy spread of IoT in businesses, homes, social settings and other kinds of environments require equivalent management and coordination speed. These areas are evolving and still needs a lot of improvement. Another notable challenge is of efficient use of energy by the IoT-enabled devices. Most of the sensors depend upon batteries as their sole source of energy. Efficient use of energy and the battery technologies have also improved in the last decade and still further research in these areas is required. Fig. 2 demonstrates the importance of IoT in smart cities.

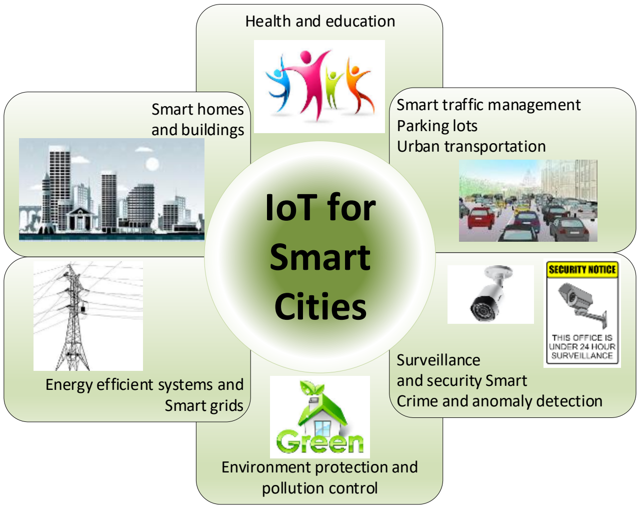

Health and education
Smart homes and buildings
Smart traffic management
Parking lots
Urban transportation
IoT for Smart Cities
Energy efficient systems and Smart grids
SECURITY NOTICE
THIS OFFICE IS UNDER 24 HOUR SURVEILLANCE
Surveillance and security Smart
Crime and anomaly detection
Green
Environment protection and pollution control

Fig. 2. The Role of IoT in Smart Cities.

F. Personal Management with IoT-Enabled Wearables

The success of IoT acceptance does not imply on the national level alone, rather, it is greatly influenced by the acceptance at the individual levels. One such example of the benefits of IoT in the personal lives of individuals is Amazon Echo, which is a voice-based interface for physically impaired people. It allows them to control several IoT devices in the home and perform daily activities with more ease.

There are a myriad of applications for personal management and monitoring comprising different IoT-enabled devices and sensors. One major application is in the health area. Smart health wearables keep a track of vital health related measurements including the heart rate, blood pressure, glucose levels etc. Special sensors and designed devices help people with different types of impairments, such as visual, cognitive, auditory or mobility impairments, in their daily activities. Fitness bands helps in monitoring a number of activities, such as the distance travelled, the number of steps taken, and the calories burnt etc. in a specified length of time.

Other prevalent use of IoT at the personal level is in the form of digital assistants. Devices and systems are capable of performing a number of assistance level activities in daily chores of individuals. Some of the examples include Amazon Alexa, Google Home, Siri, Wink and SmartThings [29]. These are Artificial Intelligence (AI) assisted products that are capable of working on voice commands and take care of trivial and advanced tasks from switching on/off the lights to scheduled maintenance of the house and taking care of the personal schedules.

Agile technology companies around the globe have started developing their products in a way that implicitly supports IoT way of life. One of the pertinent examples is the Apple's strategy and gambit that their phone work best when paired with any of the other Apple's wearable devices such as their smart watch [30]. In a recent Consumer News and Business Channel (CNBC) survey of 2017, the average American household owns at least two or more Apple products [31]. Allusive Market moves such as these paves the path to citizens tacitly embracing the IoT technology and blending it into their daily lives. Thus, contributing to the perceived and actual usefulness of IoT technology at the personal levels.

Besides all the aforementioned advantages, there a number of challenges with the use of IoT at the personal level. The first and the far most matter of concern is the security and reliability of IoT-enabled devices. Account lockout, weak passwords and illegitimate attacks mounted on poorly encrypted data pose a lot of threat to the personal data. Lack of standards play a vital role in this and efforts are being made to unify these.

III. IoT Acceptance Enablers and Challenges

Based on the discussion presented in the previous sections, we highlight, in this section, the key challenges faced by IoT adoption and present some of the main enablers of IoT acceptance that can overcome and mitigate some of the hindrances. Table I shows some of these challenges. The primary issues include:

- Security: Since the IoT technologies are still in infancy, they are subjected to a number of hostile attacks. These attacks could be targeted illegitimate, or by mistake as well, such as a device failure network scan-based random attacks. Such threats in the security of IoT devices and networks could be very dangerous in some applications such as medical, traffic control and home applications. The security threats in IoT can be classified into three categories, namely, (i) perception level (device level), (ii) transition level (network level), and (iii) application level (data level).
- Privacy: Privacy is also a major concern and contributes a lot towards the trust factor in IoT technologies along with the security factor. Privacy is further at risk when IoT devices communicate in open networks and premises are shared among residents. Research on de-anonymization techniques in the context of IoT can be very helpful.
- Trust: One of the most difficult hurdle to introduce a new technology is the trust factor of the masses. A trustworthy and reliable security and privacy mechanism is needed that is readily embraced and poses no infringements in the privacy and security of individuals. Also, there is a lack of trust when the technology has the capability of replacing humans and taking the full control over the systems.
- Scalability: The current state of the art technologies in IoT needs to incorporate the element of scalability in order to be implemented on a national scale. For this, IoT-enabled device manufacturers and service providers need to improve the quality of services, increase marketing speed, and reduce development costs. Better and new models for communication with low power consumption need to be developed instead of the client-server model existing in the most communication networks today. In addition, interoperability of devices is also very necessary in order to achieve scalability.
- Cost: In order to make it widely adopted and acceptable as a scalable technology, the cost of the IoT based devices and network components and sensors need to be minimized. The associated cost of the IoT adoption is anticipated to be less as compared to the expected gain. A higher cost could lead to hindrance in the acceptance of IoT.
- Standardization: IoT devices are manufactured by different vendors and there is a lack of common standards, especially when it comes to the communication protocol. The communication technologies between these devices also vary, for example, Blue Tooth, WiFi, IEEE 802.15.4, etc. This also makes the security among devices difficult to be maintained.
- Data Collection and Storage: With IoT technologies in practice, the amount of data is expected to explode beyond any leaps and bounds. Not only the storage and collection of such a large amount of data is an issue, the heterogeneity of the data poses greater challenge. Technologies like cloud computing and fog computing have greatly contributed to providing IoT technologies with the ability to cope with such large amounts of data.

Table II presents some of the enablers of IoT acceptance that are highly beneficial in the adoption of IoT on a national as well as personal scale in KSA. Among these the first and the far most important factor is related to the governance of IoT technology by laying down the basic rules and principles governing the IoT implementations [32]. The second important aspect is following acceptable standards for IoT implementation and continuous monitoring and analysis of these standards in collaboration with all the stakeholders. These stake holders can be from the public or private sectors. Therefore, the partnership between the public and private sectors also plays a very vital role in IoT adoption and the government, with the help of the private sector, can initiate joint projects.

TABLE I. IoT Acceptance Challenges in KSA

IoT Challenges	**Description**
Security	IoT-enabled devices and networks are still subjected to a number of legitimate and illegitimate attacks
Privacy	Devices communicating over open network pose a large threat to privacy of individuals and organizations
Trust	Lack of security, privacy and the capability of devices taking over the system leads to distrust
Scalability	Interoperability, speed in marketing and low cost is needed to achieve scalability in IoT technologies.
Cost	Cost of the current devices, sensors and the network components needs to be reduced
Standardization	Different vendors have their own standards, operating systems, hardware and communication protocols. The need for developing standards is demanded.
Data Collection and Storage	The amount of data generated with so many devices and sensors is huge. Proper mechanisms are needed to handle, store and analyze this huge data.

TABLE II. ENABLERS OF IOT TECHNOLOGY ACCEPTANCE IN KSA

	IoT Enabler	**Description**
1	Governance	Laying down the basic rules and principles governing the IoT implementations
2	Standardization	Continuous monitoring and analysis of IoT standardization process.
3	Public-Private Partnership	Joint projects can be initiated by the government with the help of the private sector
4	Awareness campaigns for the masses	A comprehensive campaign for the awareness of the masses, including the industrial workers and home users.
5	Initialization of pilot projects	Pilot projects must be initialized to pave path towards stepwise adoption and innovation in the existing technologies according to the societal needs
6	Catalyst for economic growth and better society	IoT can be considered as an important factor for the economic growth and evolution of the society.
7	Continuous Research and Development	Continuous research and development initiatives and support should be provided for implementing and developing IoT technologies enhancements
8	Protection of privacy	Protection of personal and organizational privacy should be ensured and appropriate amendments and enhancements in the law should be made when necessary
9	Right to disconnect	The users of the IoT networks should have the right to be disconnected and conceal their identities.
10	Emerging risks identification and mitigation	Emerging risks should be identified and countermeasures should be established to deal with the surfacing risks to privacy

For the awareness of IoT in the masses, a comprehensive campaign is needed that includes meetings with the industrial workers and home users. Another enabler is the initiative to initialize pilot projects at different levels. Pilot projects pave path towards stepwise adoption and innovation in the existing technologies according to the societal needs. Thus, enabling IoT as a catalyst for the economic growth and the evolution of the society. Therefore, Continuous research and development initiatives and support should be provided for implementing and developing IoT technologies enhancements.

Protection of privacy of individuals and organizations is another enabler of IoT adoption. Privacy protection should be ensured and appropriate amendments and enhancements in the law should be made when necessary. The users of the IoT networks should have the right to be disconnected and conceal their identities. In addition, Emerging risks should be identified and countermeasures should be established to deal with the surfacing risks to privacy.

IV. CONCLUSION

This work provides an insight to the primary determinants of the acceptance of IoT in the KSA. It highlights several application areas at the national as well as the individual levels. These include the acceptance in the industry, agriculture and livestock, education, health, smart cities, personal management with IoT enabled wearables, and AI assisted technologies. Along with the potential benefits that contribute to the acceptance and perceived acceptance of the technology, there are a number of risks that are identified. These challenges include security and privacy of IoT, trust, cost of devices and components, scalability, standardization, and data collection and storage issues. IoT can be a way of life where its proper acceptance require mitigation of these issues. Lastly, this work highlights some enablers that could contribute to the swift acceptance of IoT technologies in KSA.

ACKNOWLEDGMENT

The authors highly acknowledge the financial support provided by the Deanship of Scientific Research (DSR) at Saudi Electronic University, Saudi Arabia to publish this research.

REFERENCES

[1] H. Lee, “Home IoT resistance: Extended privacy and vulnerability perspective,” Telematics and Informatics, vol. 49, p. 101377, Jun. 2020, doi: 10.1016/j.tele.2020.101377.

[2] S. Kim and S. Kim, “User preference for an IoT healthcare application for lifestyle disease management,” Telecommunications Policy, vol. 42, no. 4, pp. 304–314, May 2018, doi: 10.1016/j.telpol.2017.03.006.

[3] P. Brous, M. Janssen, and P. Herder, “The dual effects of the Internet of Things (IoT): A systematic review of the benefits and risks of IoT adoption by organizations,” International Journal of Information Management, vol. 51, p. 101952, Apr. 2020, doi: 10.1016/j.ijinfomgt .2019.05.008.

[4] R. Pillai and B. Sivathanu, “Adoption of internet of things (IoT) in the agriculture industry deploying the BRT framework,” Benchmarking: An International Journal, vol. 27, no. 4, pp. 1341–1368, Jun. 2020, doi: 10.1108/BIJ-08-2019-0361.

[5] T. Sudtasan and H. Mitomo, “The Internet of Things as an accelerator of advancement of broadband networks: A case of Thailand,” Telecommunications Policy, vol. 42, no. 4, pp. 293–303, May 2018, doi: 10.1016/j.telpol.2017.08.008.

[6] J. Robert, S. Kubler, and S. Ghatpande, “Enhanced Lightning Network (off-chain)-based micropayment in IoT ecosystems,” Future Generation Computer Systems, vol. 112, pp. 283–296, Nov. 2020, doi: 10.1016/j.future.2020.05.033.

[7] S. Teixeira et al., “LAURA architecture: Towards a simpler way of building situation-aware and business-aware IoT applications,” Journal of Systems and Software, vol. 161, p. 110494, Mar. 2020, doi: 10.1016/j.jss.2019.110494.

[8] “Forecast: Internet of Things — Endpoints and Associated Services, Worldwide, 2017,” Gartner. https://www.gartner.com/en/documents/ 3840665/forecast-internet-of-things-endpoints-and-associated-ser (accessed Aug. 26, 2020).

[9] A. Sestino, M. I. Prete, L. Piper, and G. Guido, “Internet of Things and Big Data as enablers for business digitalization strategies,” Technovation, Jan. 2020, doi: 10.1016/j.technovation.2020.102173.

[10] M. S. Farooq, S. Riaz, A. Abid, K. Abid, and M. A. Naeem, “A Survey on the Role of IoT in Agriculture for the Implementation of Smart Farming,” IEEE Access, vol. 7, pp. 156237–156271, 2019, doi: 10.1109/ACCESS.2019.2949703.

[11] Z. Khan, A. Anjum, K. Soomro, and M. A. Tahir, “Towards cloud based big data analytics for smart future cities,” J Cloud Comp, vol. 4, no. 1, p. 2, Feb. 2015, doi: 10.1186/s13677-015-0026-8.

[12] I. Usman and K. A. Almejalli, “Intelligent Automated Detection of Microaneurysms in Fundus Images Using Feature-Set Tuning,” IEEE Access, vol. 8, pp. 65187–65196, 2020, doi: 10.1109/ACCESS. 2020.2985543.

[13] A. Albesher, “IoT in Health-care: Recent Advances in the Development of Smart Cyber-Physical Ubiquitous Environments,” IJCNS, vol. 19, no. 2, pp. 181–186.

[14] S. M. R. Islam, D. Kwak, MD. H. Kabir, M. Hossain, and K.-S. Kwak, “The Internet of Things for Health Care: A Comprehensive Survey,”

IEEE Access, vol. 3, pp. 678–708, 2015, doi: 10.1109/ACCESS.2015.2437951.

[15] A. Majeed and M. Ali, "How Internet-of-Things (IoT) making the university campuses smart? QA higher education (QAHE) perspective," in 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), Jan. 2018, pp. 646–648, doi: 10.1109/CCWC.2018.8301774.

[16] "The cheap security cameras inviting hackers into your home – Which? News." https://www.which.co.uk/news/2019/10/the-cheap-security-cameras-inviting-hackers-into-your-home/ (accessed Aug. 17, 2020).

[17] T. Naheed, I. Usman, and A. Dar, "Lossless data hiding using optimized interpolation error expansion," in Frontiers of Information Technology (FIT), 2011, 2011, pp. 281–286.

[18] S. Hu, B. Hu, and Y. Cao, "The wider, the better? The interaction between the IoT diffusion and online retailers' decisions," Physica A: Statistical Mechanics and its Applications, vol. 509, pp. 196–209, Nov. 2018, doi: 10.1016/j.physa.2018.06.008.

[19] Jieh-Haur CHEN, Nguyen Thi Thu HA, Hsing-Wei TAI, and Chao-An Chang, "The Willingness to Adopt the Internet of Things (IoT) Conception in Taiwan's Construction Industry," Journal of Civil Engineering & Management, vol. 26, no. 6, pp. 534–550, Aug. 2020, doi: 10.3846/jcem.2020.12639.

[20] R. Gupta, K. Miyazaki, and Y. Kajikawa, "Ingredients of Successful Emerging Business Ecosystems: Case of Industrial IoT Adoption," 2018 Portland International Conference on Management of Engineering and Technology (PICMET), 2018, doi: 10.23919/PICMET.2018.8481854.

[21] R. Gorli, "A New Approach for Employee Safety in Industries with IoT," i-Manager's Journal on Information Technology; Nagercoil, vol. 7, no. 2, pp. 22–29, May 2018, doi: http://dx.doi.org.sdl.idm.oclc.org/10.26634/jit.7.2.14650.

[22] M. Younan, E. H. Houssein, M. Elhoseny, and A. A. Ali, "Challenges and recommended technologies for the industrial internet of things: A comprehensive review," Measurement, vol. 151, p. 107198, Feb. 2020, doi: 10.1016/j.measurement.2019.107198.

[23] I. Khan, I. Usman, T. Usman, G. U. Rehman, and A. U. Rehman, "Intelligent churn prediction for telecommunication industry," International Journal of Innovation and Applied Studies, vol. 4, no. 1, pp. 165–170, 2013.

[24] M. Ayaz, M. Ammad-Uddin, Z. Sharif, A. Mansour, and E.-H. M. Aggoune, "Internet-of-Things (IoT)-Based Smart Agriculture: Toward Making the Fields Talk," IEEE Access, vol. 7, pp. 129551–129583, 2019, doi: 10.1109/ACCESS.2019.2932609.

[25] F. A. Majid and N. M. Shamsudin, "Identifying Factors Affecting Acceptance of Virtual Reality in Classrooms Based on Technology Acceptance Model (TAM)," Asian Journal of University Education, vol. 15, no. 2, pp. 51–60, Dec. 2019.

[26] H. Shaikh, M. S. Khan, Z. A. Mahar, M. Anwar, A. Raza, and A. Shah, "A Conceptual Framework for Determining Acceptance of Internet of Things (IoT) in Higher Education Institutions of Pakistan," in 2019 International Conference on Information Science and Communication Technology (ICISCT), Mar. 2019, pp. 1–5, doi: 10.1109/CISCT.2019.8777431.

[27] H. Rajab and T. Cinkelr, "IoT based Smart Cities," in 2018 International Symposium on Networks, Computers and Communications (ISNCC), Jun. 2018, pp. 1–4, doi: 10.1109/ISNCC.2018.8530997.

[28] R. Bukhsh, N. Javaid, M. I. Khan, Z. A. Khan, and I. Usman, "Cost efficient hybrid techniques for DSM in smart homes," International Journal of Ad Hoc and Ubiquitous Computing, vol. 33, no. 2, pp. 90–108, Jan. 2020, doi: 10.1504/IJAHUC.2020.105462.

[29] B. Caddy, N. Pino, and H. S. L. 3 days ago, "The best smart speakers 2020: which one should you buy?," TechRadar Middle East. https://www.techradar.com/news/best-smart-speakers (accessed Sep. 28, 2020).

[30] D. Gershgorn, "The Apple Watch Is the New Starter Phone," Medium, Sep. 15, 2020. https://onezero.medium.com/the-apple-watch-is-the-new-starter-phone-4eb74ed61e0b (accessed Sep. 16, 2020).

[31] S. Liesman, "America loves its Apple. Poll finds that the average household owns more than two Apple products," CNBC, Oct. 10, 2017. https://www.cnbc.com/2017/10/09/the-average-american-household-owns-more-than-two-apple-products.html (accessed Sep. 20, 2020).

[32] Yonghee Kim, Youngju Park, and Gwangsuk Song, "Interpretive Structural Modeling in the Adoption of IoT Services," KSII Transactions on Internet & Information Systems, vol. 13, no. 3, pp. 1184–1198, Mar. 2019, doi: 10.3837/tiis.2019.03.004.

[33] E. Shaikh and N. Mohammad, "Applications of Blockchain Technology for Smart Cities," in 2020 Fourth International Conference on Inventive Systems and Control (ICISC), Jan. 2020, pp. 186–191, doi: 10.1109/ICISC47916.2020.9171089.

[34] H. Tranter, "A survey on approaches to the protection of personal data gathered by IoT devices," PeerJ PrePrints; San Diego, Jul. 2018, doi: http://dx.doi.org.sdl.idm.oclc.org/10.7287/peerj.preprints.26473v2.

Fraud Detection in Credit Cards using Logistic Regression

Hala Z Alenzi[1], Nojood O Aljehane[2]
Department of Computer Science
Tabuk University, Tabuk City
Kingdom Saudi Arabia

Abstract—**Due to the increasing number of customers as well as the increasing number of companies that use credit cards for ending financial transactions, the number of fraud cases has increased dramatically. Dealing with noisy and imbalanced data, as well as with outliers, has accentuated this problem. In this work, fraud detection using artificial intelligence is proposed. The proposed system uses logistic regression to build the classifier to prevent frauds in credit card transactions. To handle dirty data and to ensure a high degree of detection accuracy, a pre-processing step is used. The pre-processing step uses two novel main methods to clean the data: the mean-based method and the clustering-based method. Compared to two well-known classifiers, the support vector machine classifier and voting classifier, the proposed classifier shows better results in terms of accuracy, sensitivity, and error rate.**

Keywords—Classifier; logistic regression; accuracy; smoothing; artificial intelligence; cross validation

I. Introduction

According to the definition of fraud [1], the aim of fraud is to achieve personal or financial gain through deception. Based on this, fraud detection and prevention are the two significant methods for avoiding loss due to fraud. Fraud prevention is the proactive technique for avoiding the occurrence of fraudulent acts, and fraud detection is the technique for the detection of fraudulent transactions by fraudsters [2]. A variety of payment cards, including credit, charge, debit, and prepaid cards, are currently widely available. They are the most popular means of payment in some countries [3]. Indeed, advances in digital technologies have paved the way for changes in how we handle money, especially for payment methods that have changed from being a physical activity to a digital activity using electronics means [4]. This has revolutionized the landscape of monetary policy, including the business strategies and operations of both large and small companies. Credit card fraud is the fraudulent use of credit card details to buy a product or service. These transactions can be physically or digitally performed [5]. In physical transactions, the credit card is physically present. On the other hand, digital transactions take place over the internet or telephone. A cardholder normally provides their card number, card verification number, and expiration date through a website or telephone call. With the rapid rise in e-commerce over the past few years, credit card use has increased tremendously [1,3].

In Malaysia, the number of transactions performed through credit cards in 2011 was approximately 317 million, and this number increased to 447 million in 2018 [4]. In 2015, global credit card fraud reached a record of $21.84 billion, as reported by [2]. The number of fraud cases has been rising with the increased use of credit cards. While various verification methods have been implemented, the number of fraud cases involving credit cards has not been significantly decreased [6]. The potential for substantial monetary gains, combined with the ever-changing nature of financial services, creates a wide range of opportunities for fraudsters [7]. Funds from payment card fraud are often used in criminal activities that are hard to prevent, e.g., to support terrorist acts [8]. The internet is where fraudsters prefer to be because they are able to conceal their location and identity. The recent increase in credit card fraud has directly hit the financial sector hard. Losses due to credit card fraud mainly impact merchants because they bear all expenses, including the fees from their card issuer, administrative fees and other charges [9]. All the losses are borne by the merchants, leading to increases in the prices of goods and decreases in discounts. Hence, reducing this loss is highly important. An effective fraud detection system is required to minimize the number of cases of fraud.

A. Motivation

The use of credit cards to perform financial transactions at banks or other institutions is a common action in light of the currently available technology. Online payments (or any other online transactions) bring benefits to companies and individuals in terms of the convenience, velocity, and flexibility of performing daily duties [10,11]. The work in [12] presented a statistical analysis related to the usage of credit cards over five years (from 2006 to 2010). This reflected the huge dependency on credit cards by both people and organizations. To take advantage of advanced technologies, companies try to use advanced techniques to provide high-quality services to customers. Automation can be seen as the best solution for attracting more customers and consequently collecting more financial gain [13]. The process of converting a manual system to a fully automatic on, as found in smart cities, is not without risk.

B. Problem Statement

According to [14], it is estimated that 10,000 transactions take place via credit cards every second worldwide. Owing to such a high transaction frequency, credit cards have become the primary targets of fraud. Indeed, since the Diners Club released its first credit card in 1950, credit card companies have been fighting against fraud [15]. Every year, billions of dollars are lost directly because of credit card fraud. Fraud cases occur under different conditions, e.g., transactions at points of sale (POSs) or transactions made online or over the telephone, i.e., card-not-present (CNP) cases or transactions with lost and stolen cards. In this way, the credit card fraud in 2015 alone amounted to $21.84 billion, with issuers bearing $15.72 billion of the cost [16]. Based on information from the European Central Bank, in 2012, the majority (60%) of fraud stemmed from CNP transactions, and another 23% stemmed from POS terminals. The value of fraud is high globally and locally in Malaysia. The volumes of credit, debit, and charge cards were 383.8 million, 107.6 million, and 4.1 million, respectively, in 2016 and increased to 447.1 million, 245.7 million, and 5.2 million, respectively, in 2018 [9]. The overall percentage of fraudulent payments (i.e., with credit, debit, and charge cards) was 0.0186% in 2016 and increased by 37.6% to 0.0256% in 2018 [17]. The potential for huge monetary gains combined with the ever-changing nature of financial services provides opportunities for fraudsters. In Malaysia, 1,000 card transactions occur every minute. Fraud directly impacts merchants and financial institutions because they incur all the costs. An increase in fraud affects customers' confidence in using electronic payments [18].

Many surveys have shown that the increase in the dependence on credit cards to perform financial transactions is accompanied by an increasing rate of fraud, as seen in [1,3]. The increasing capabilities of the attackers or the hackers have accentuated the problem since these people can exploit security gaps to obtain sensitive information about users or their credit information to perform malicious activities, such as fraud [4,5]. To define this problem accurately, Fig. 1 shows the general scenario of performing credit card fraud.

As shown in Fig. 1, the attacker can perform malicious activities on many sides of the online process. To solve this problem, a fraud detection system is needed. Artificial intelligence (AI) is defined as the research field that aims at performing machine learning to obtain an intelligent machine that can perform tasks on behalf of the user. This can be done through two main steps: training and testing. AI is employed to build systems for fraud detection, such as classification-based systems [19,6,7,8], clustering-based systems [17,20,21], neural network-based systems [18,22,23], and support vector machine-based systems [9]. Although AI-based systems can perform well, they suffer from some critical issues. First, the term "imbalanced data" refers to unbalanced data used for training, where one class of the data is dominated by the other (i.e., the majority of data belong to one class and the rest belong to the other). This negatively affects the accuracy of detection [24,25]. Second, the term "noisy data" refers to the existence of outliers within the data employed for training. Outliers can be seen outside of the normal context of the data. This issue also leads to poor detection accuracy [26,16]. Third, the concept of drift means that the behaviour of the client changes, resulting in changes in the data stream when dealing with online data detection in real time [15,14].

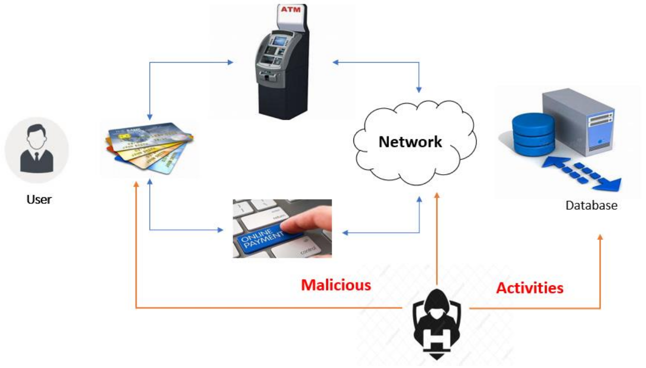


Fig. 1. General Scenario of Online Fraud.

C. Research Questions

On the basis of the empirical evidence, the following research questions are developed to guide this study and meet its objectives.

- How can a fraud detection system be built using AI that can deal with imbalanced data effectively?
- How can we smooth (or clean) the data before using it for training the machine to ensure high detection accuracy?
- How can the system detect fraud by adapting to the behaviour of the user?

D. Contributions

The contributions of this work can be summarized as follows:

- An AI-based system for fraud detection is proposed. The system uses logistic regression to build a classifier called the LogR classifier. The LogR classifier has the ability to deal with imbalanced data and adapt to the behaviour of the user by employing the cross-validation technique.
- To ensure high accuracy detection, two main methods are used to clean the data. The mean-based method deals with missing values, and the clustering-based method deals with outliers.
- Extensive experiments are conducted to train and test the proposed classifier using a standard database.

E. Structure of the Paper

The rest of this work is organized as follows. Section II reviews the related work. Section III describes the proposed artificial intelligence system in detail. In Section IV, the metrics used are presented for evaluation purposes. Section V presents the experiments and discusses the results in light of a comparison with similar approaches. Finally, the paper is concluded in Section VI.

II. Related Work

This section first provides a brief background about the research domain. Then, the related work is presented in detail.

A. Background

The background refers to the credit card research field in terms of the intersection of multiple research sectors. This field can be viewed as the intersection of four main domains, as illustrated in Fig. 2.

The definitions of the domains and terms that are applied in this study are listed below.

Artificial Intelligence (AI): It can be defined as the science that addresses the methods used for training machines to mimic the brains of humans. In other words, machines can be used to make decisions on behalf of human users. In this context, data mining tasks, such as classification, clustering, applying association rules, and using neural networks, are employed [2].

Financial Systems: These can be defined as the systems that are used to convert manual transactions into digital transactions. In this context, the term “transaction” denotes any financial activity that may be performed by a user based on a specific system [27].

Chip Industry: This term refers to the manufacturing of chips to store critical information on the card of the user. The information acts as a key to trigger any transaction. However, the chip is programmed to match some passwords to allow access to financial interfaces [28].

Internet of Things (IoT): It can be defined as a collection of devices connected via a network. The devices vary from small devices with low processing power (such as watches) to large devices high processing power, such as mobile devices. Using IoT devices to perform financial transactions is vital, especially in light of the goal of shifting toward smart cities [29].

B. Groups of AI-based Techniques

Artificial intelligence (AI) is defined as enabling machines to make decisions on behalf of human users. In this context, data mining tasks, such as classification, clustering, applying association rules, and using neural networks, are employed [2]. In addition, AI is employed to build systems for fraud detection, such as classification-based systems [19,6,7,8], clustering-based systems [17,20,21], neural network-based systems [18,22,23] and support vector machine-based systems [9].

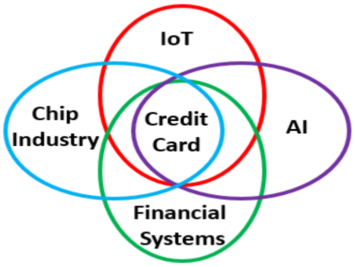


Fig. 2. The Intersection of Credit Card Research and other Research Fields.

The techniques employed to construct credit card fraud detection systems using AI can be categorized into four main groups. This idea is shown in Fig. 3.

1) Classification-based systems: The authors in [19] tried to achieve two main objectives in their work: (1) enhancing the accuracy of the classifications output by credit card fraud detection systems and (2) lowering the response times of these systems. To achieve the first goal, the authors proposed a hybrid model that fuses two classifiers to generate a new (or enhanced) one. The first classifier used is the K-means classifier, which deals with overlapping data because such data cause poor accuracy. The second classifier is the artificial bee colony algorithm (ABC), which is used to enhance the performance of the system. The first classifier forms the first level, and the second classifier forms the second level of the classification process proposed in the same model. The database used in this work was generated by using the C# programming language, where the number of instances was 100,000. In addition, 12 features were selected to include in the training phase. The selected features were based on a rule engine.

Moreover, previous systems suffered from problems in real-time environments [6]. These are problems in the context of credit card fraud detection. Such problems include imbalanced data, noisy data, and the concept of drift. The authors applied the bag creation technique to solve the data problems; this technique involves performing the sampling process on the collected data in real time. To clean the data, they applied naïve Bayes networks for the effective manipulation of noisy data. An incremental learning-based method was presented to address the concept of drift. The data set used in this work is summarized in Table I.

The strength of this study is the enhancement in performance achieved by using Spark to implement the system in parallel. In addition, the reduction in cost is considered an important feature of this system, and this was achieved by employing naive Bayes networks in the process of classification. The weakness of the proposed system is that it does not manipulate cyclic recurrences that may be included in the concept of drift. Cyclic recurrences refer to cyclic repetitions in the distributions of data.

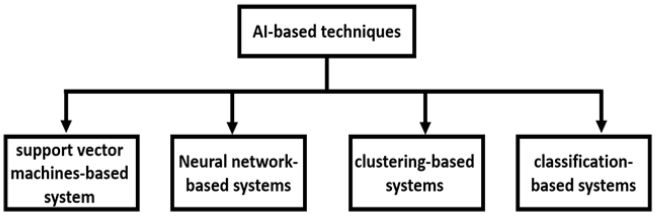

AI-based techniques
support vector machines-based system
Neural network-based systems
clustering-based systems
classification-based systems

Fig. 3. Categories of AI-based Techniques for Fraud Detection.

TABLE I. USED DATA SET [6]

Start day	End day	Instances	%Fraudulent Transactions
July 2004	September 2004	0.3 million	3.74%

The authors in [7] evaluated the current fraud detection system with regard to credit card transactions. The problem is that there are two stages for automatic classification: real-time (RT) and near-real-time (NRT). They focused on the NRT stage by using a rule-based classification technique that considers the final evaluation of the human element of fraud. The authors did not improve the design of the system, discover any new rules, or improve the arithmetic efficiency of individual rules. Instead, they manipulated the rules to form a decision-making system to improve both the accuracy and the performance. The key idea is to calculate the contribution of each rule involved in the system. Calculating the contribution of a rule depends on the difference between two values, which are (1) the performance of the system when the rule is used and (2) the performance of the system without using the rule. The degree of performance improvement is high if the rule is not redundant and is low if it is redundant with other rules or rule groups. For the measurement of performance, the precision, recall and F-score metrics were employed. A real database, which consists of 359,862 records provided by some industrial partners, was used for the training phase.

The authors in [8] addressed credit card fraud detection. In this study, the authors relied on the fact that "the features of the financial transactions in institutions change over time". This shows that the problem of credit card fraud detection should be considered in real time. Therefore, they converted this problem into real working transactions. In terms of artificial intelligence, the class should not be provided to the classifier immediately during the training stage. The key idea of the proposed approach is to follow a strict strategy that has three main steps: (1) analysing the real conditions under which the real transactions are performed; (2) employing these conditions to train the classifier using two main data sets; and (3) testing the classifier after the training stage is completed and supporting it by using the feedback of the users (their interactions) to improve the accuracy of the classifier. Table II summarizes the dataset used.

2) Clustering-based systems: To address the problem of detecting credit card fraud through transactions, the authors in [17] dealt with the problem of online shopping fraud and the concept of drift. They proposed a strategy consisting of four stages: (1) based on both the previous transaction data and the information of the cardholders, they used the clustering method to divide the cardholders into different groups for the purpose comparing their behaviours; (2) they proposed a sliding window strategy to group the transactions in each group to extract the behavioural patterns for each cardholder; (3) they trained a set of classifications for each group to measure behavioural patterns; and (4) they used a group of classifiers by training them on cardholder behaviours and output the highest behaviour pattern. A feedback mechanism was used to solve the concept of drift problem. Four dataset simulators were generated to manually create the data sets.

The authors in [20] proposed a clustering-based method. In this study, the fraud detection problem in ecommerce is manipulated and may be exploited by hackers who are highly skilled. The methods proposed to address such problems suffer from low accuracy and effectiveness. In addition, the methods used for detecting fraud may make some mistakes in identifying fraudulent transactions. The reason behind such shortcomings is that the proposed approaches focus on order analysis rather than anything else. Motivated by these facts, the authors proposed a method that focuses on the hackers themselves. The key idea is to extract some recognized features, such as the address of delivery, customer name, and methods of payment, and then, based on these features, the similarity among the attackers is calculated. Based on these similarities, the attackers are grouped in some clusters for detection. A main feature of their proposed method is that two current methods, agglomerative clustering and sampling, are selectively used in a reasonable amount of time for recursively grouping orders into small clusters. The dataset used for the training process was inspired by the Zalando website. This website periodically receives approximately 29 million orders (some of them are normal and others are fraudulent).

The authors in [21] tried to evaluate the detection problem by extracting the general pattern of the dataset to represent the fraud. In other words, the enhancement of the clustering methods relies only on the clusters used; this technique is called general enhancement. The authors proposed an approach that enables the application of local enhancement as well as general enhancement for fraud detection in financial transactions. They proposed the “Hierarchical Clusters-based Deep Neural Networks (HC-DNN)” method that uses the anomalous features of hierarchical clusters that are pretrained based on an autoencoder as the initial weights for neural networks. In detail, the data are grouped based on abnormal features that refer to fraud. These features are then used as the initial weights for the input layers of neural networks, as shown in Fig. 4.

TABLE II. DETAILS OF THE DATASET USED [8]

Id	**End day**	**Instances**	**Features**	**%Fraudulent Transactions**
2013	2014-01-18	21'830'330	51	0.19%
2014-2015	2015-05-31	54'764'384	51	0.24%

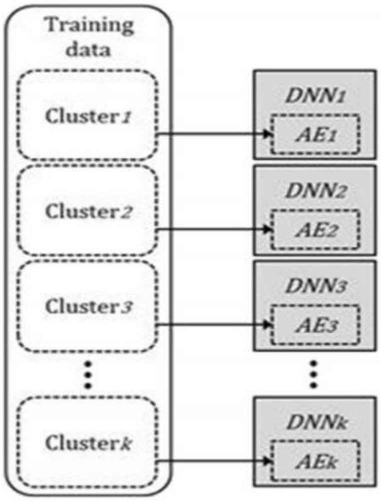


Fig. 4. Key Idea of the HC-DNN Method [21].

The authors used a dataset containing 19,505 records, including fraudulent and non-fraudulent records. The dataset is skewed and consists of 19,313 non-fraudulent and 192 fraudulent cases. Some preprocessing steps were performed on the data to mitigate the negative impact of the imbalanced data before using them for actual training.

3) Neural network-based systems: The authors in [18] discussed issues related to increasing fraud detection in online shopping transactions and payments, especially those related to credit cards. To detect credit card fraud, they proposed a neural network-based system. It uses back prorogation to enhance the output of the neural network so that the error (the difference between the actual or desirable value and the output of the neural network) is distributed back by adjusting the weights of the inputs. The strategy followed in this work can be summarized through the following steps:

a) A new Neuroph Project was created in Neuroph Studio using the Java programming language.

b) The actual perceptron network was constructed.

c) The training data set was prepared.

d) The training process was started by considering the desired value (the accuracy of fraud detection) set by an expert in the field.

e) The trained network was tested.

The data used for training were collected from a data mining blog. It includes 20000 active credit card holders with transactions spanning more than six months. The authors in [22] proposed a “Convectional Neural Network CNN” in their work. Similar to previous works, the problem studied was how to detect a pattern that represents fraudulent transactions. In their method, the CNN forms a classifier that takes features of the transactions as inputs. The features are extracted from each transaction and stored in a feature matrix. The classifier has the ability to deal with imbalanced data based on the sampling technique. The key idea behind the sampling technique is to use higher than normal costs to generate fraudulent transactions. Fig. 5 illustrates the general scenario of the CNN model.

The data used includes more than 260 million credit card transactions in one year. Approximately four thousand transfers are listed as fraudulent, and the remainder are legal. A hybrid fraud detection system was proposed in [23]. The key idea is to use neural networks as classifiers. Since the network needs to update the weights of the input layer, a swarm optimization method was employed for this purpose. Finally, the model was tested and evaluated. Fig. 6 illustrates the general structure of the proposed system, which is called the “Particle Swarm Optimization Auto-associative Neural Network (PSOAANN)”.

4) Support vector machine-based systems: The authors in [9] used a support vector machine (SVM) to improve the accuracy of the classifier in the process of detecting fraud in credit card transactions. The key idea behind using an SVM is to split the features that represent transactions, where these features are used for the clustering process. In other words, the data are cleaned initially. Then, the features of transactions are extracted. Third, the features are measured to calculate the similarity among them. To isolate the features as much as possible, the SVM is used. Fourth, the K-means clustering algorithm is used to cluster the data based on the isolated (i.e., as far as possible) features. The classifier is then trained on the clusters. The classifier that deals with fraudulent transactions is used to detect fraud. The database used for training contains 5310 records in total. Among them, 490 records are fraudulent data and 4820 are non-fraudulent data, and 1174 characteristic variables are included.

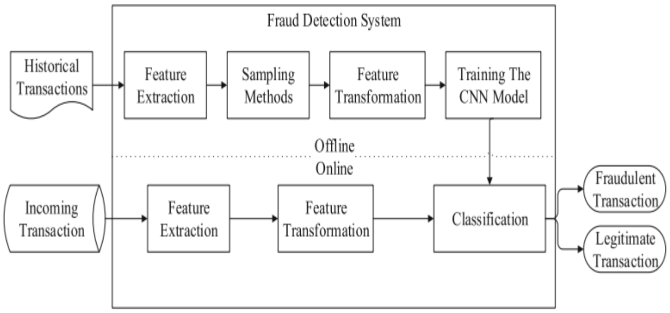


Fig. 5. General Scenario of the Fraud Detection System Proposed in the Work in [22].

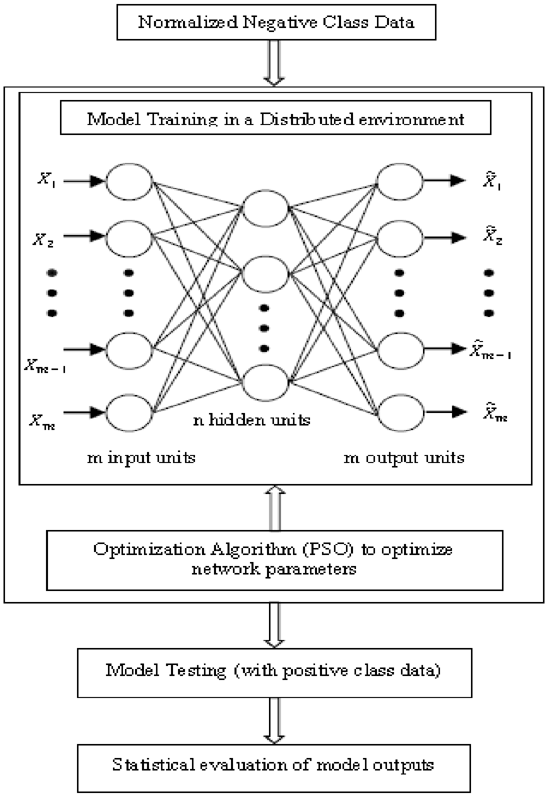

Normalized Negative Class Data
Model Training in a Distributed environment
X_1
X_2
X_{m-1}
X_m
n hidden units
m input units
m output units
$\hat{X}_1$
$\hat{X}_2$
$\hat{X}_{m-1}$
$\hat{X}_m$
Optimization Algorithm (PSO) to optimize network parameters
Model Testing (with positive class data)
Statistical evaluation of model outputs

Fig. 6. Structure of the PSOAANN-based System [23].

III. Proposed Approach

This section describes the proposed approach in detail. Fig. 7 illustrates the steps of the proposed approach.

As shown in Fig. 7 above, there are nine steps, starting with the selection of the database and ending with the use of the classifier in real-life situations. The reason behind selecting logistic regression to build the classifier is related to its efficiency of detecting frauds based on its ability to isolate the data that belong to different binary classes.

A. Selecting the Database

This work uses a standard dataset that is available on the internet [30]. The dataset contains transactions made using credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred over two days, where we have 492 fraudulent cases out of 284,807 transactions. The dataset is highly unbalanced, and the positive class (fraudulent cases) accounts for 0.172% of all transactions. Fig. 8 shows the selection step in the implemented programme represented by “Load DB”.

As shown in Fig. 8, the loading of the data is competed, and the size of the dataset can be seen.

To explore the data contained in this data set, Fig. 9 shows the data exploration options that can be chosen.

As shown in Fig. 9, there are 6 views of the used data set. This enables us to clearly explore the database. In terms of exploring the database, Fig. 10 and 11 show two examples of data exploration.

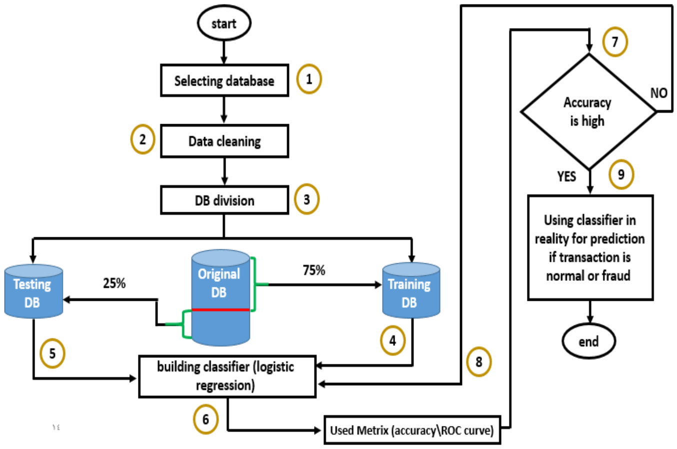
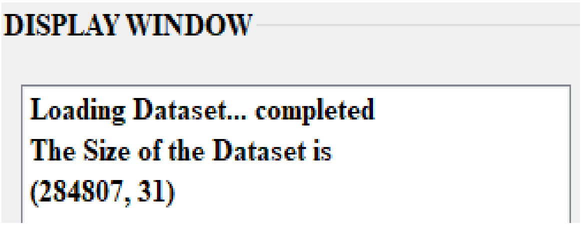

start
Selecting database
1
2
Data cleaning
DB division
3
Testing DB
25%
Original DB
75%
Training DB
4
5
building classifier (logistic regression)
8
6
Used Metrix (accuracy\ROC curve)
7
Accuracy is high
NO
YES
9
Using classifier in reality for prediction if transaction is normal or fraud
end

Fig. 7. Flow Chart of the Proposed Approach.



DISPLAY WINDOW

Loading Dataset... completed
The Size of the Dataset is
(284807, 31)

Fig. 8. Loading the used Dataset.

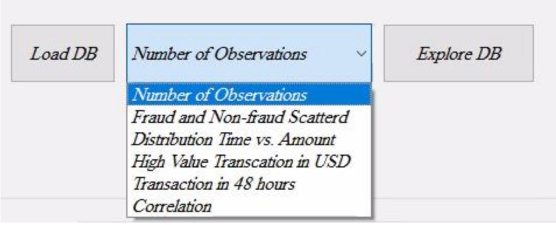


Fig. 9. Interface for Selecting (or Loading) the Data Set.

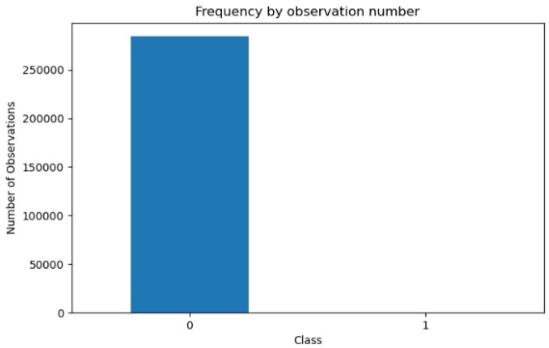


Fig. 10. Data Exploration based on the Observation Number.

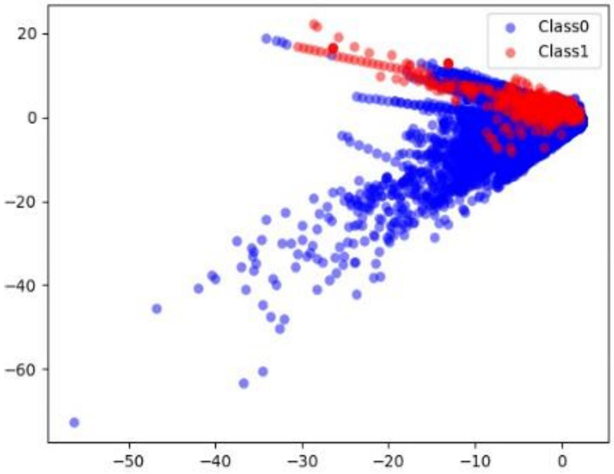


Fig. 11. Data Exploration based on the Two Main Classes of the Data.

B. Data Cleaning

The goal of this step is to clean the data and prepare it for the training phase of the classifier. In general, data in reality are noisy. Therefore, a cleaning step is necessary. In the context of the data cleaning process, the procedure is as follows:

1) Fill in the missing values. A missing value means that a cell of a given record is empty due to an mistake during entry.

2) Solve any inconsistencies. This means that if there is a collision in the data, this collision must be resolved.

3) Remove any outliers. Outliers refer to abnormal values (i.e., very high values or very low values).

Fortunately, most of the data used in the data set are cleaned except for some missing values and outliers. The mechanism that is used for handling the missing values depends on the mean (mathematical operation) since the data are numbers. Fig. 12 illustrates to the process of filling in a missing value.

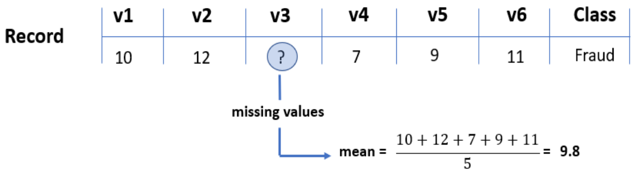
For the handling of outliers, a clustering-based method is employed in this work. The key idea is to create three clusters (one for the normal data, a second one for high values, and a third for low values). After grouping the data into the clusters, the last two clusters (i.e., those that contain outliers) are deleted. Fig. 13 illustrates the mechanism of outlier removal.

C. Database Division

In this step, the database is divided into training and testing databases. The goal of the training database is to construct the classifier (model), while the goal of the testing database is to test (evaluate) the built classifier. In this work, the cross-validation method is used to divide the database, which is divided into 10 parts, as shown in Fig. 14.

As shown in Fig. 14, the database is divided into 10 parts (i.e., the value of $k = 10$ in the cross-validation method). In the first iteration ($k = 1$), the first nine parts are considered a training set, while the last part of the database is considered a testing set. In the second iteration ($k = 2$), both the first eight parts and the tenth part are considered as a training set, while the ninth part of the database is considered a testing set. This process continues until the last iteration ($k = 10$), where the first part is the testing set and the last nine parts are the training set.

Fig. 15 illustrates a sample of the code execution process based on the cross-validation method when clicking on the “Split DB” button.



Record	v1	v2	v3	v4	v5	v6	Class
	10	12	?	7	9	11	Fraud

$$\text{mean} = \frac{10+12+7+9+11}{5} = 9.8$$

Fig. 12. Mean-based Mechanism for Handling Missing Values.

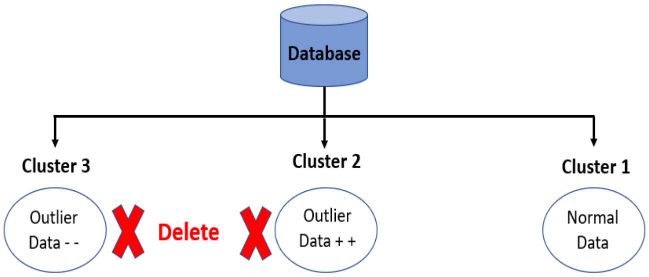


Fig. 13. Mechanism of Outlier Removal.

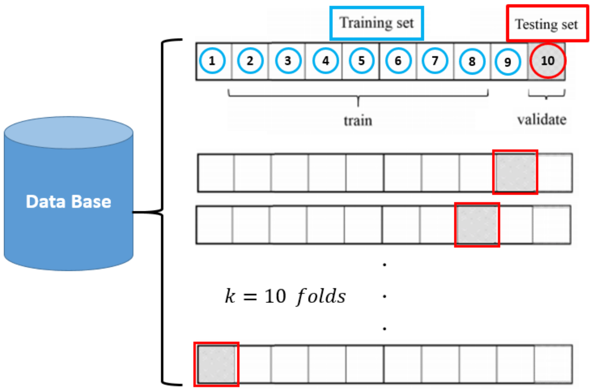
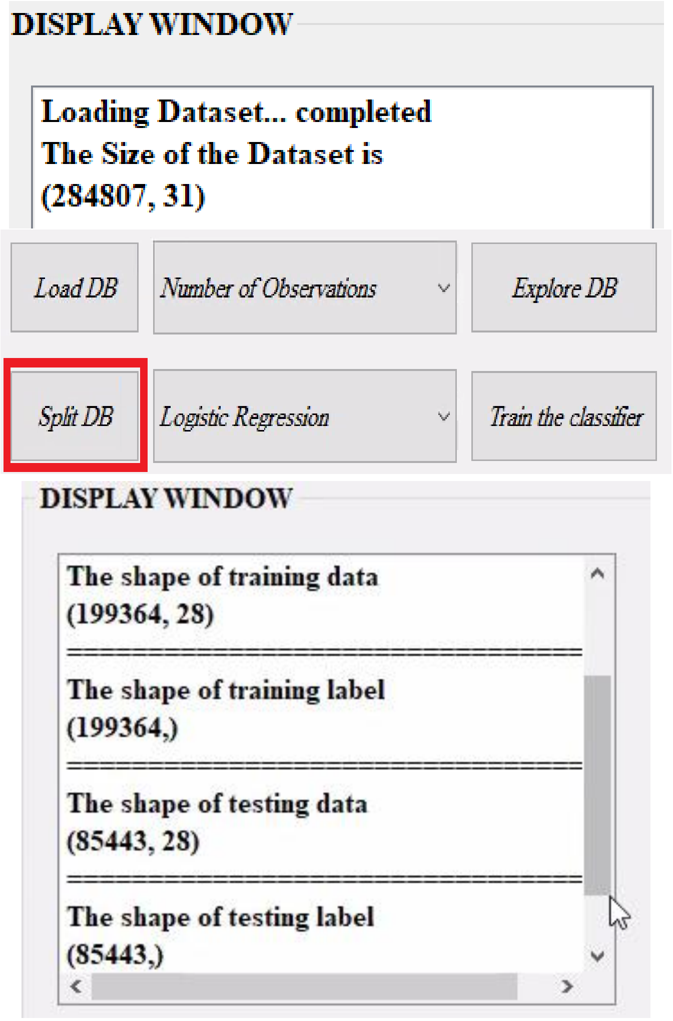


Fig. 14. Division of the Database based on Cross Validation.



DISPLAY WINDOW

Loading Dataset... completed
The Size of the Dataset is
(284807, 31)

DISPLAY WINDOW

The shape of training data
(199364, 28)
=====================================
The shape of training label
(199364,)
=====================================
The shape of testing data
(85443, 28)
=====================================
The shape of testing label
(85443,)

Fig. 15. Results of the Division Process.

D. Building the Classifier

In the context of building the classifier, logistic regression is employed. Logistic regression is more advanced than linear regression. The reason for this is that linear regression cannot classify data that are widely distributed in a given space, as shown in Fig. 16.

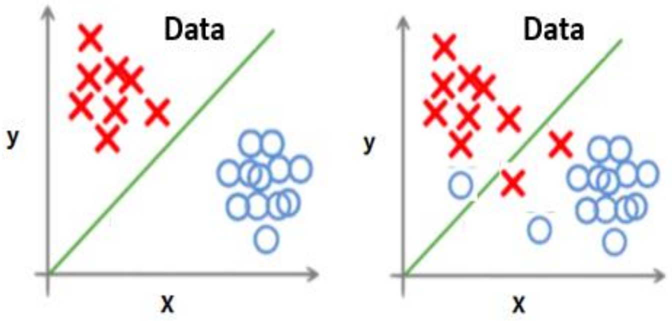


Fig. 16. The Limitation of Linear Regression.

As shown in Fig. 16, on the left side, the linear regression has the ability to classify the data, where the line can divide the given data into two main categories (or classes). The right side of Fig. 16 illustrates the limitation of linear regression. When the data overlap, the line cannot divide the data into two clear classes. This limitation is overcome by logistic regression. Fig. 17 provides a visual comparison between the linear regression and the logistic regression methods for the purpose of highlighting this limitation.

Logistic regression has the following advantages [32]:

1) Logistic regression is easier to implement than linear regression and is very efficient to train.

2) It makes no assumptions about the distributions of classes in the feature space.

3) It can easily be extended to multiple classes (multinomial regression).

4) It is very efficient for classifying unknown records.

The logistic regression equation can be obtained from the linear regression equation. The mathematical steps to obtain logistic regression equations are given below:

The equation of the straight line can be written as:

$$y = a_0 + a_1 \times x_1 + a_2 \times x_2 + \cdots a_k \times x_k \quad (1)$$

In logistic regression, y can be between 0 and 1 only, so we divide the above equation by $(1 - y)$:

$$\frac{y}{1-y} \; |0 \; for \; y = 0 \; and \; \infty \; for \; y = 1 \quad (2)$$

As a result, the logistic regression equation is defined as:

$$\log \; [\frac{y}{1-y}] = a_0 + a_1 \times x_1 + a_2 \times x_2 + \cdots a_k \times x_k \quad (3)$$

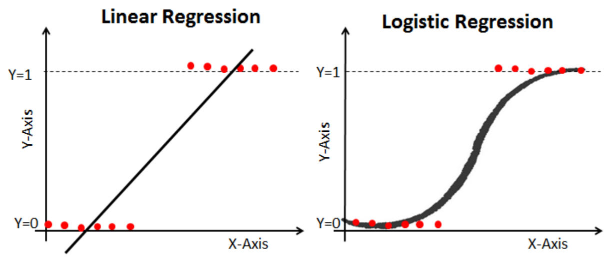


Fig. 17. A Visual Comparison between Linear and Logistic Regression [31].

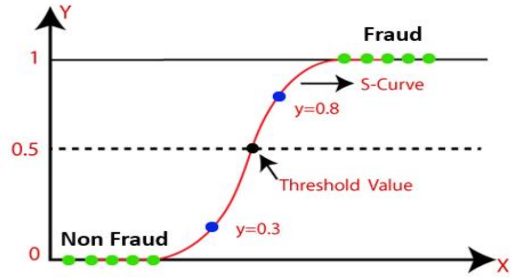


Fig. 18. The Concept of Logistic Regression Classification [33].

In other words, the fraud class takes the value "1", while the non-fraud class takes the value "0". A threshold of 0.5 is used to differentiate between the two classes, as shown in Fig. 18.

E. Testing the Classifier

Since the cross-validation method divides the database into 10 parts, there are 10 testing data sets. Each testing data set is used to test one classifier (there are 10 classifiers). This in turn gives the model an advantage by allowing it to use the whole database for testing as well as for training. The testing process is tightly coupled with the accuracy of the model. Calculating the final accuracy involves calculating the accuracy of each classifier. Formally, let Acc_k^C denote the accuracy of a given trained classifier, as shown in Fig. 19.

Then, the final accuracy of the final classifier (ACC_F^C) is obtained based on the "average" mathematical operation.

$$ACC_F^C = \frac{\sum_{k=1}^{10} Acc_k^C}{k} \quad (4)$$

F. Evaluating the Classifier

In general, a confusion matrix is an effective benchmark for analysing how well a classifier can recognize records of different classes [34]. The confusion matrix is formed based on the following terms:

1) True positives (TP): positive records that are correctly labelled by the classifier.

2) True negatives (TN): negative records that are correctly labelled by the classifier.

3) False positives (FP): negative records that are incorrectly labelled positive.

4) False negatives (FN): positive records that are mislabelled negative.

Table III shows the confusion matrix in terms of the TP, FN, FP, and TN values.

Relying on the confusion matrix, the accuracy, sensitivity, and error rate metrics are derived. For a given classifier, the accuracy can be calculated by considering the recognition rate, which is the percentage of records in the test set that are correctly classified (fraudulent or non-fraudulent). The accuracy is defined as:

$$Accuracy = \frac{(TP+TN)}{number \; of \; all \; records \; in \; the \; testing \; set} \quad (5)$$

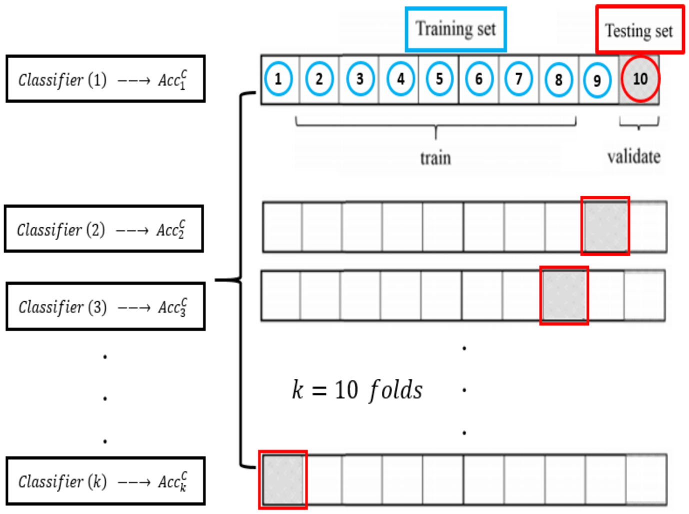

Training set
Testing set
Classifier (1) ⟶ Acc_1^C
1
2
3
4
5
6
7
8
9
10
train
validate
Classifier (2) ⟶ Acc_2^C
Classifier (3) ⟶ Acc_3^C
$k = 10\ folds$
Classifier (k) ⟶ Acc_k^C

Fig. 19. Classifiers with Corresponding Accuracies.

Mechanisms for accuracy-based evaluation. In this context, a higher accuracy corresponds to a better classifier output. The maximum value of the accuracy metric is 1 (or 100%), which is achieved when the classifier classifies the records correctly without any errors in the classification process.

Sensitivity refers to the true positive recognition rate. It is given by:

$$Sensitivity = \frac{TP}{P} \qquad (6)$$

Mechanisms for sensitivity-based evaluation. In this context, a higher sensitivity corresponds to a better classifier output. The maximum value of the sensitivity metric is 1 (or 100%), which is achieved when the proportion of true positive cases equals the number of actual positive cases.

The error rate is defined as the ratio of mistakes made by the classifier during the prediction process. It is defined as:

$$eror\ rate = 1 - accuracy \qquad (7)$$

Mechanisms for error rate-based evaluation. In this context, a higher accuracy corresponds to a worse classifier output. The maximum value of the accuracy metric is 1 (or 100%), which is achieved when the classifier classifies all the records incorrectly (i.e., the accuracy is zero).

G. Examining the Value of the Accuracy

In this step, the final calculated accuracy is examined. If it is accepted, then the classifier can be used in real-life situations. Otherwise, the process of building the classifier has a problem, and then retraining the classifier is required.

TABLE III. CONFUSION MATRIX

Actual class (Predicted class)	**Confusion matrix**		
	C1	**¬ C1**	**Total**
C1	True positives (TP)	False negatives (FN)	TP + FN = P
¬ C1	False positives (FP)	True negatives (TN)	FP + TN = N

Security and privacy issues are highly stressed according to many studies [35-43] when using data in the artificial intelligence research field. This is because the data reflect the policies and sensitive issues of the institution in question (these are banks in our work when applying the proposed classifier in reality). Therefore, the privacy and security of data are not considered in this work, but they will be considered in future work.

IV. Used Metrics

Since the domain of this work is artificial intelligence, two types of metrics are used. They are AI-based metrics and performance-based metrics.

A. AI-based Metrics

In this context, the confusion matrix dominates the situation. In other words, the metrics that are derived from the confusion matrix are employed to measure the prediction accuracy of the classifier.

B. Performance-based Metrics

In this context, time dominates the situation. In other words, the total time ($ToTi$) required to build, train, and test the classifier is used as a benchmark. The $ToTi$ is given by:

$$ToTi = T_{pre} + T_{dbs} + T_{tr} + T_{ts} \qquad (8)$$

where T_{pre} refers to the preprocessing time, T_{dbs} refers to the database splitting time, T_{tr} refers to the training time, and T_{ts} refers to the testing time. It is well known that the lower the total time is, the higher the degree of performance.

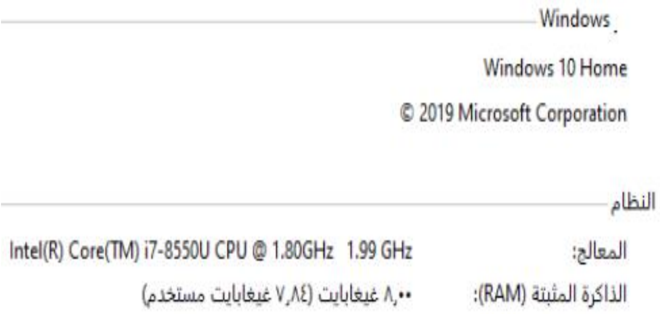
V. Results and Discussions

This section is structured so that the specifications of the machine used to implement the proposed classifier are introduced. Then, the classifiers that are compared with the proposed classifier are described. Finally, the results are provided along with two discussions.

A. Setup

The system is performed on a machine that has the specifications summarized in Fig. 20.

The programming language used for the implementation of the classifier is Python.



Windows

Windows 10 Home

© 2019 Microsoft Corporation

النظام

المعالج: Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz 1.99 GHz

الذاكرة المثبتة (RAM): ٨,٠٠ غيغابايت (٧,٨٤ غيغابايت مستخدم)

Fig. 20. Specifications of the Machine used to Implement the Classifiers.

B. Selected Classifiers

Two classifiers are selected for a comparison with the classifier proposed in this work. They are the K-nearest neighbours (KNN) classifier and the voting classifier (VC). Below, a brief description of each selected classifier is presented.

Fig. 21 shows the fundamental steps required to build the voting classifier.

As shown in Fig. 21, there are many classifiers, and a voting step is required to produce the final output class. The voting step means that the final output of the classifier depends on the majority of the classes (predictions) that are generated by the classifiers. For example, there are three classifiers in Fig. 22. The final prediction is either Fraud (F) or Non-Fraud (NF). The voting process works as follows:

1) Obtain the outputs of the classifiers.

2) Calculate the number of classifiers that generate the F class (let us say 2 classifiers).

3) Calculate the number of classifiers that generate the NF class (let us say 1 classifier).

4) The majority is 2. Therefore, the final prediction is the F class.

Fig. 22 shows the fundamentals steps for building the KNN classifier.

As shown in Fig. 22, there are two clusters (one for fraudulent transactions and one for non-fraudulent transactions). Each cluster has a centre, which is represented numerally by (-1) for nonfraudulent transactions and (+1) for fraudulent transactions. For a given transaction, the KNN classifier processes the transaction and generates a corresponding number. Then, the distance between the generated value and the centre of each cluster is calculated. Finally, the transaction is assigned to the correct cluster (in the example, it is assigned to the non-fraud cluster).

C. Results

Since the cross-validation method is used to divide the database, we obtain ten sub-classifiers as mentioned previously. The process of calculating the final values of the AI-based metrics depends on the "average" mathematical operation. Table IV summarizes the obtained results.

Table V summarizes the comparison of the logistic regression (LogR)-based classifier with both the KNN-based classifier and the VC-based classifier.

Discussion. From Table V, it is obvious that the LogR classifier achieves the best values in terms of accuracy, sensitivity, and error rate. The reason behind this is related to the efficient preprocessing technique used to remove outliers and manipulate the missing values. In addition, cross validation ensures that the entire database is employed as both the training and testing data sets, and this in turn enhances the three metrics. The KNN classifier comes in second, and the VC classifier comes in third. This is because the KNN classifier includes a step related to calculating the distances between the value of the new transaction and the centres of clusters. This in turn reflects efficient processing in the prediction process compared to poor processing in the VC classifier (i.e., only calculating the majority).

For performance comparison purposes, the bare chart shown in Fig. 23 illustrates the values of the response time for all classifiers involved in the comparison.

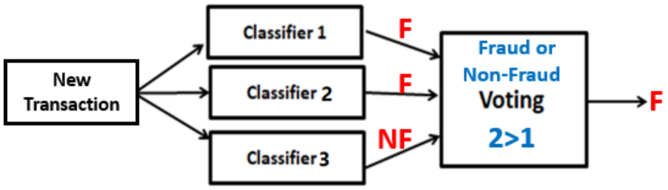


Fig. 21. Basic Concept of the Voting Classifier.

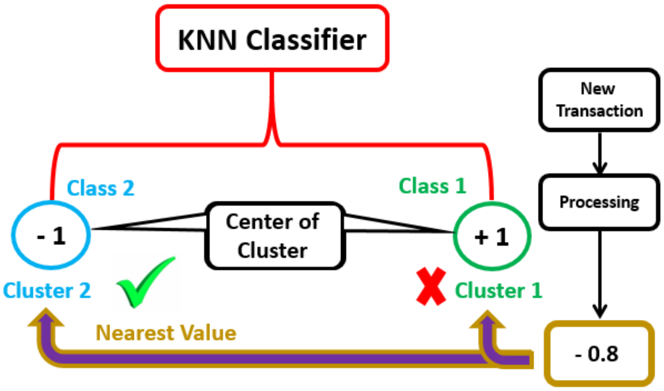


Fig. 22. Basic Concept of the KNN Classifier.

TABLE IV. EVALUATING THE PROPOSED CLASSIFIER

K-value	Accuracy	Sensitivity	Error rate
1	96%	97%	4%
2	98%	96%	2%
3	98%	97%	2%
4	96%	96%	4%
5	97%	98%	3%
6	96%	98%	4%
7	97%	96%	3%
8	98%	98%	2%
9	98%	98%	2%
10	98%	96%	2%
Average	97.2%	97%	2.8%

TABLE V. COMPARISON OF CLASSIFIERS

Classifier	Metrics		
	Accuracy	*Sensitivity*	*Error rate*
LogR classifier	97.2%	97%	2.8%
KNN classifier	93%	94%	7%
VC classifier	90%	88%	10%

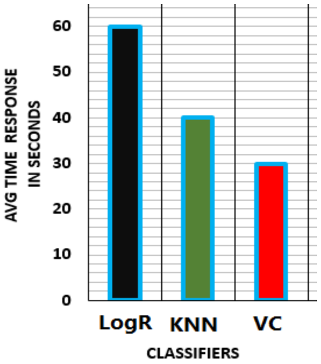


Fig. 23. Performances of the Three Classifiers.

Discussion. Fig. 23 shows that the VC classifier achieves the best performance. This is because it depends only on a simple mathematical operation (the sum operation) to determine the classes and generate the final output. The KNN classifier comes in the second in terms of its response time. That is because this classifier must perform additional mathematical operations related to calculating the distances between the new value and the centre of each cluster, and these operations in turn consumes more time. Compared to the previous classifiers, the LogR classifier performs the worst. The reason for this is that the time required for database division and training the sub-classifiers is very high. In other words, training and testing ten sub-classifiers logically takes time less than training and testing one classifier (i.e., the KNN and VC classifiers). However, although the response time of the LogR classifier is the longest, it achieves the best accuracy. From the point of view of detecting fraud (or security), accuracy more of a concern than performance. This issue will be taken into consideration in future work.

VI. Conclusion

The detection of credit card fraud is a vital research field. This is because of the increasing number of fraud cases in financial institutions. This issue opens the door for employing artificial intelligence to build systems that can detect fraud. Building an AI-based system to detect fraud requires a database to train the system (or classifier). The data in reality are dirty and have missing values, noisy data, and outliers. Such issues negatively affect the accuracy rate of the system. To overcome these problems, a logistic regression-based classifier is proposed. The data are first cleaned using two methods: the mean-based method and clustering-based method. Second, the classifier is trained based on the cross-validation technique (folds=10), which ensures that the whole database is used as both the training data set and testing data set. Finally, the proposed classifier is evaluated based on the accuracy, sensitivity, and error rate metrics. The proposed logistic regression-based classifier is compared to well-known classifiers, which are the K-nearest neighbours classifier and the voting classifier. The logistic regression-based classifier generates the best results (accuracy = 97.2%, sensitivity = 97%, and error rate = 2.8%).

Limitations. The performance of the proposed classifier suffers in terms of response time. In addition, it does not apply to data in real time.

Future work. In future work, we intend to enhance the performance and take the security and privacy of the data in real time into consideration.

References

[1] Yousefi, Niloofar, Marie Alaghband, and Ivan Garibay. "A Comprehensive Survey on Machine Learning Techniques and User Authentication Approaches for Credit Card Fraud Detection." arXiv preprint arXiv:1912.02629 (2019).

[2] Paschen, Jeannette, Jan Kietzmann, and Tim Christian Kietzmann. "Artificial intelligence (AI) and its implications for market knowledge in B2B marketing." Journal of Business & Industrial Marketing (2019).

[3] Abdallah, Aisha, Mohd Aizaini Maarof, and Anazida Zainal. "Fraud detection system: A survey." Journal of Network and Computer Applications 68 (2016): 90-113.

[4] Alladi, Tejasvi, et al. "Consumer IoT: Security vulnerability case studies and solutions." IEEE Consumer Electronics Magazine 9.2 (2020): 17-25.

[5] Rahman, Rizwan Ur, et al. "Classification of Spamming Attacks to Blogging Websites and Their Security Techniques." Encyclopedia of Criminal Activities and the Deep Web. IGI Global, 2020. 864-880.

[6] Somasundaram, Akila, and Srinivasulu Reddy. "Parallel and incremental credit card fraud detection model to handle concept drift and data imbalance." Neural Computing and Applications 31.1 (2019): 3-14.

[7] Gianini, Gabriele, et al. "Managing a pool of rules for credit card fraud detection by a Game Theory based approach." Future Generation Computer Systems 102 (2020): 549-561.

[8] Dal Pozzolo, Andrea, et al. "Credit card fraud detection: a realistic modeling and a novel learning strategy." IEEE transactions on neural networks and learning systems 29.8 (2017): 3784-3797.

[9] Wang, Chunhua, and Dong Han. "Credit card fraud forecasting model based on clustering analysis and integrated support vector machine." Cluster Computing 22.6 (2019): 13861-13866.

[10] Deufel, Patrick, Jan Kemper, and Malte Brettel. "Pay now or pay later: A cross-cultural perspective on online payments." Journal of Electronic Commerce Research 20.3 (2019): 141-154.

[11] Co-Pending, U. S. "patent application No." US201514696366, filed on Apr 24 (2015).

[12] Hamid, N. R., and Aw Yoke Cheng. "A risk perception analysis on the use of electronic payment systems by young adult." WSEAS Transactions on Information Science and applications 10.1 (2013): 26-35.

[13] inc website (2020), online available : https://www.inc.com/guides/cust_tech/20909.html, access (10 March 2020).

[14] Janbandhu, Ruchika, Shameedha Begum, and N. Ramasubramanian. "Credit Card Fraud Detection." Computing in Engineering and Technology. Springer, Singapore, 2020. 225-238.

[15] Mittal, Sangeeta, and Shivani Tyagi. "Computational Techniques for Real-Time Credit Card Fraud Detection." Handbook of Computer Networks and Cyber Security. Springer, Cham, 2020. 653-681.

[16] Zou, Junyi, Jinliang Zhang, and Ping Jiang. "Credit Card Fraud Detection Using Autoencoder Neural Network." arXiv preprint arXiv:1908.11553 (2019).

[17] Jiang, Changjun, et al. "Credit card fraud detection: A novel approach using aggregation strategy and feedback mechanism." IEEE Internet of Things Journal 5.5 (2018): 3637-3647.

[18] Murli, Divya, et al. "Credit card fraud detection using neural networks." International Journal of Students' Research in Technology & Management 2.2 (2015): 84-88.

[19] Darwish, Saad M. "An intelligent credit card fraud detection approach based on semantic fusion of two classifiers." Soft Computing 24.2 (2020): 1243-1253.

[20] Marchal, Samuel, and Sebastian Szyller. "Detecting organized eCommerce fraud using scalable categorical clustering." Proceedings of the 35th Annual Computer Security Applications Conference. 2019.

[21] Kim, Jeongrae, Han-Joon Kim, and Hyoungrae Kim. "Fraud detection for job placement using hierarchical clusters-based deep neural networks." Applied Intelligence 49.8 (2019): 2842-2861.

[22] Fu, Kang, et al. "Credit card fraud detection using convolutional neural networks." International Conference on Neural Information Processing. Springer, Cham, 2016.

[23] Kamaruddin, Sk, and Vadlamani Ravi. "Credit card fraud detection using big data analytics: use of PSOAANN based one-class classification." Proceedings of the International Conference on Informatics and Analytics. 2016.

[24] Arun, C., and C. Lakshmi. "Class Imbalance in Software Fault Prediction Data Set." Artificial Intelligence and Evolutionary Computations in Engineering Systems. Springer, Singapore, 2020. 745-757.

[25] Thabtah, Fadi, et al. "Data imbalance in classification: Experimental evaluation." Information Sciences 513 (2020): 429-441.

[26] Maung, Ei Thinzar Win. Comparison of Data Mining Classification Algorithms: C5. 0 and CART for Car Evaluation and Credit Card Information Datasets. Diss. Unversity of Computer Studies, Yangon, 2020.

[27] Rike, James B. "Cylinder support system." U.S. Patent Application No. 29/641,843.

[28] Freund, Peter C. "Method and system for performing purchase and other transactions using tokens with multiple chips." U.S. Patent No. 10,282,536. 7 May 2019.

[29] Benamar, Lamya, Christine Balagué, and Zeling Zhong. "Internet of Things devices appropriation process: the Dynamic Interactions Value Appropriation (DIVA) framework." Technovation 89 (2020): 102082.

[30] Kaggle , website (2020). Avaliable : https://www.kaggle.com/mlg-ulb/creditcardfraud (access 22 July 2020).

[31] DataCamp , website (2020). Avaliable : https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python (access 28 July 2020).

[32] Salillari, Denisa, and Luela Prifti. "Comparison Study of Logistic Regression Model for Albanian Texts." *Journal of Advances in Mathematics* 12.9 (2016): 6572-6575.

[33] javatpoint , website (2020). Avaliable : https://www.javatpoint.com/logistic-regression-in-machine-learning (access 22 July 2020).

[34] Mona Alfifi, Mohamad Shady Alrahhal, Samir Bataineh and Mohammad Mezher, “Enhanced Artificial Intelligence System for Diagnosing and Predicting Breast Cancer using Deep Learning” International Journal of Advanced Computer Science and Applications(IJACSA), 11(7), 2020. http://dx.doi.org/10.14569/IJACSA.2020.0110763

[35] Alrahhal, Mohamad Shady, Maher Khemakhem, and Kamal Jambi. "Agent-Based System for Efficient kNN Query Processing with Comprehensive Privacy Protection." International Journal Of Advanced Computer Science And ApplicationS 9.1 (2018): 52-66.

[36] 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.

[37] Alrahhal, Mohamad Shady, Maher Khemakhem, and Kamal Jambi. "A SURVEY ON PRIVACY OF LOCATION-BASED SERVICES: CLASSIFICATION, INFERENCE ATTACKS, AND CHALLENGES." Journal of Theoretical & Applied Information Technology 95.24 (2017).

[38] Alrahhal, Mohamad Shady, Maher Khemekhem, and Kamal Jambi. "Achieving load balancing between privacy protection level and power consumption in location based services." (2018).

[39] Alrahhal, H.; Alrahhal, M.S.; Jamous, R.; Jambi, K. A Symbiotic Relationship Based Leader Approach for Privacy Protection in Location Based Services. ISPRS Int. J. Geo-Inf. 2020, 9, 408.

[40] Al-Rahal, M. Shady, Adnan 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): 29.

[41] Alluhaybi, Bandar, et al. "A Survey: Agent-based Software Technology Under the Eyes of Cyber Security, Security Controls, Attacks and Challenges." International Journal of Advanced Computer Science and Applications (IJACSA) 10.8 (2019).

[42] Fouz, Fadi, et al. "Optimizing Communication And Cooling Costs In Hpc Data Center." Journal of Theoretical and Applied Information Technology 85.2 (2016): 112.

[43] Alrahhal, Mohamad Shady, and Adnan Abi Sen. "Data mining, big data, and artificial intelligence: An overview, challenges, and research questions." (2018).

New Sector Scan Geometry for High Frame Rate 2D-Echocardiography using Phased Arrays

Wided Hechkel[1], Néjib Hassen[3]
Laboratory uEI
University of Monastir
Monastir, Tunisia

Brahim Maaref[2]
Laboratory EuE
University of Monastir
Monastir, Tunisia

***Abstract*—2D echocardiography high frame rate techniques do have some drawbacks such as crosstalk artifacts caused by the interactions between the parallel transmitted and received beams. In this paper, we suggest a new cardiac imaging technique based on MLT (Multi-Line Transmission). The main idea of our approach is to benefit from the scan geometry to reduce the interference between the simultaneously transmitted beams. We propose to do the scan at different depths and in parallel to the diagonal scan sector. Therefore, compared to existing MLT techniques, the new scan sector strategy will result in artifacts' reduction in the ultrasound imaging systems. We entitled our approach the Synthetic Sum of Multi-Line Transmission (SS-MLT). Simulations of Point Spread Function (PSF), multiple Point Spread Functions (PSFs) and Cyst Phantom (CP) provided in this paper are compared in our approach to the main classical ultrasound imaging approaches. Therefore, the SS-MLT exhibits a very similar lateral profile as the Single Line Transmission (SLT) algorithm. Hence, the simulation results indicate a potential value of a future hardware implementation of SS-MLT technique.**

***Keywords*—2D Echocardiography; high frame rate; multi-line transmit beamforming; new scan geometry; reduced crosstalk**

I. Introduction

The most critical challenge with the 2D echocardiography is to operate at a high temporal resolution without adversely affecting the image quality of the system. A higher frame rate may allow a more precise recognition of the overall cardiac morphology and its mechanical events. In diagnostic ultrasound imaging systems, the frame rate relies essentially on three parameters; speed of sound over the tissue, penetration depth and the number of events' transmission per frame.

Many researches were about increasing the frame rate by decreasing the number of transmit events for one frame. Among these researches we mention Diverging Wave (DW) imaging [1,2], Multiline Acquisition (MLA) [3,4] and Multiline Transmission (MLT) [5-9]. All these methods operate at an increasing temporal resolution compared to the conventional cardiac imaging techniques. They tried to come up with solutions that give a good tradeoff between image quality and high frame rate. But all the ultrafast techniques mentioned above relatively cause a loss in Signal to Noise Ratio (SNR).

The most used method is the MLT which involves transmitting multiple ultrasonic beams simultaneously with a gain in frame rate similar to the amount of parallel emissions [10–19]. The fundamental concept beyond the MLT technique is a parallel transmission of several focused pulses in different directions, separated by a static opening angle.

But, the problem of this method is the crosstalk artifacts generated by interference on transmission and reception [8] - [13]–[18]. Many researches focused on solving these problems, one possibly with the use of changed frequency ranges for every sent pulse [19]–[20], the adoption of the approaches namely pulse inversion to erase the influences of the extra insonified trends [21] or control the windowing weights of the emitted and acquired signals in order to conform the orientation of the main lobes of the concerned waves, to the orientation of the zeros in the sidelobe fields of the extra waves [22].In [23],the given authors propose a simultaneous multi-zone focusing method using orthogonal quadratic chirp signals to improve the lateral resolution without sacrificing the frame rate. In the proposed method, two weighted quadratic chirp signals with different spectra are simultaneously transmitted with different transmission time delays for a multi-zone focusing. Because the two weighted quadratic chirps can be designed to have a desired level of cross-correlation after compression, the degradation of axial resolution resulting from the division of a spectrum is minimized. In [24], the given authors suggest using the Second-Harmonic Signal in MLT. Therefore, taking advantage of the nonlinear propagation of sound within the tissue, the second-harmonic signal can be used with the MLT technique. The image obtained using the second-harmonic signal, compared to an image obtained by using the fundamental signal, should have reduced artifacts coming from other pulses transmitted simultaneously. In [25] coded excitation has been put forward for crosstalk suppression.

Inappropriately, no one of these techniques has shown an effective suppression of the crosstalk artifacts without affecting the spatial resolution of the output image. Also, most of them exhibit a high circuitry complexity and require a very difficult implementation.

We bring to the fore a new solution based on the geometries that take advantage of three main mechanisms. First, our method uses the MLT approach to form 9 LRIs (Low Resolution Images) and combines them to reconstruct one HRI (High Resolution Image). Second, it adopts a sparse aperture in transmission to avoid interference between adjacent beams, where each of the apertures is connected to an appropriate pulse generator. Third, the focusing points for simultaneous transmission beams are split diagonally in parallel to the diagonal scan sector, which can dramatically decrease the crosstalks in the image. Using this approach, we maintain a frame rate of about 570 HZ allowing 2-D echocardiographic applications.

In this paper, our objective is to demonstrate that our proposed method can minimize the interference dramatically between transmission/reception pressure fields. Simulations' results of the PSF (Point Spread Function) and CP (Cyst Phantom) prove that this method reduces the production of undesirable artifacts and produces an image quality (IQ) comparable to the IQ of the DRF algorithm.

The following section explains the theroretical background of the different crosstalks in the MLT. Section III presents the new approach for a full MLT sector scan geometry for 2d echocardiography. Simulation results for PSF (Point Spread Function) and CP (Cyst Phantom) are presented in Section IV. Section V discusses SS-MLT technique results in comparison with other algorithms. The inferences and conclusions along with the future scope of study is presented in Section VI.

II. Theoretical Background: MLT-MLA

MLT algorithm is subject to the crosstalk artifacts generated by the interference between beams on transmission and on reception [8] - [13] - [18]. These crosstalks can be classified into three types [13]: transmission crosstalk, reception crosstalk, and transmission/reception crosstalk.

A. Transmission Crosstalk

In this section, we explain how and why the transmission crosstalks between all the ultrasound pressure fields appear in the system. Fig. 1 illustrates the overall transmission crosstalk process in the case of 8-MLT where eight simultaneous beams' focus are at the same depth, but at different angles. The given figure highlights the interactions between the fourth and fifth regions' targets. These communications are the same for all adjacent transmission directions.

We notice an overlapping between the side lobes of the transmission line and the main lobe of the adjacent reception lines steered in a changed orientation. Hence, the side lobe energy of the first transmission beam is picked up by the main lobe of the second reception beam. The number of crosstalks is important under these circumstances.

Transmission crosstalk generates a perturbation that affects measured data with spurious patterns. It is due to the energy transmitted outside the steering direction. The principle three causes of transmission crosstalk are as follows: first the small angular separation between the simultaneously transmitted beams. Second, the less directive beam with high sidelobe levels is also responsible for increasing those perturbations. Third, the same focusing depth shared between all parallel transmission lines is a dramatic factor that produces crosstalks, and this is explained whenever an array element receives a signal that is not reflected from its intended target but generated by neighbour elements and picked up with its own generated signals at the same time.

B. Reception Crosstalk

Fig. 2 illustrates the reception crosstalks in the system. The ultrasound pressure fields mentioned in the figure are measured under the same conditions as transmission crosstalk; focusing at the same depth with 8MLT and each line is insonified from one sub-aperture. We can easily remark an overlapping between the main lobe of the transmission line and the side lobes of the adjacent reception lines steered in a changed orientation. Hence, the main lobe energy of the first transmission beam is picked up by the side lobe of the second reception beam. Also, here, the number of crosstalks is significantly important.

Reception crosstalk produces an interference leading to a high noise in the detected patterns. This noise is due to the effect of echoes coming from scatterers that are not situated in the direction of interest. Like transmission crosstalk, the central causes of reception crosstalk are the small angular separation between the parallel transmitted lines, the ignore of using an apodization to the aperture at reception and, applying the same focusing depth to all parallel transmit directions that generates conjunction in time between transmit and receive fields.

C. Transmision / Reception Crosstalk

Fig. 3 demonstrates the transmission/reception crosstalks in the system. The ultrasound pressure fields mentioned in the aforementioned figure are taken under the same depth with 8MLT and each line is insonified from one sub-aperture. We notice an overlapping between the side lobes of the transmission line and the side lobes of the adjacent reception lines steered in a changed orientation. Hence, the side lobe energy of the first transmission beam is picked up by the side lobe of the second reception beam. Hence, the crosstalks' density is increased under these conditions appear like parasitic blocks in the ultrasound image. This type of crosstalks has the same causes of appearing like transmission crosstalk and reception crosstalk. Hence, the non-adoption of apodization windows in transmission and reception, the thin separating angle between transmission directions, both with the conjunction in time between reception and transmission signals all help this crosstalks' type to appear.

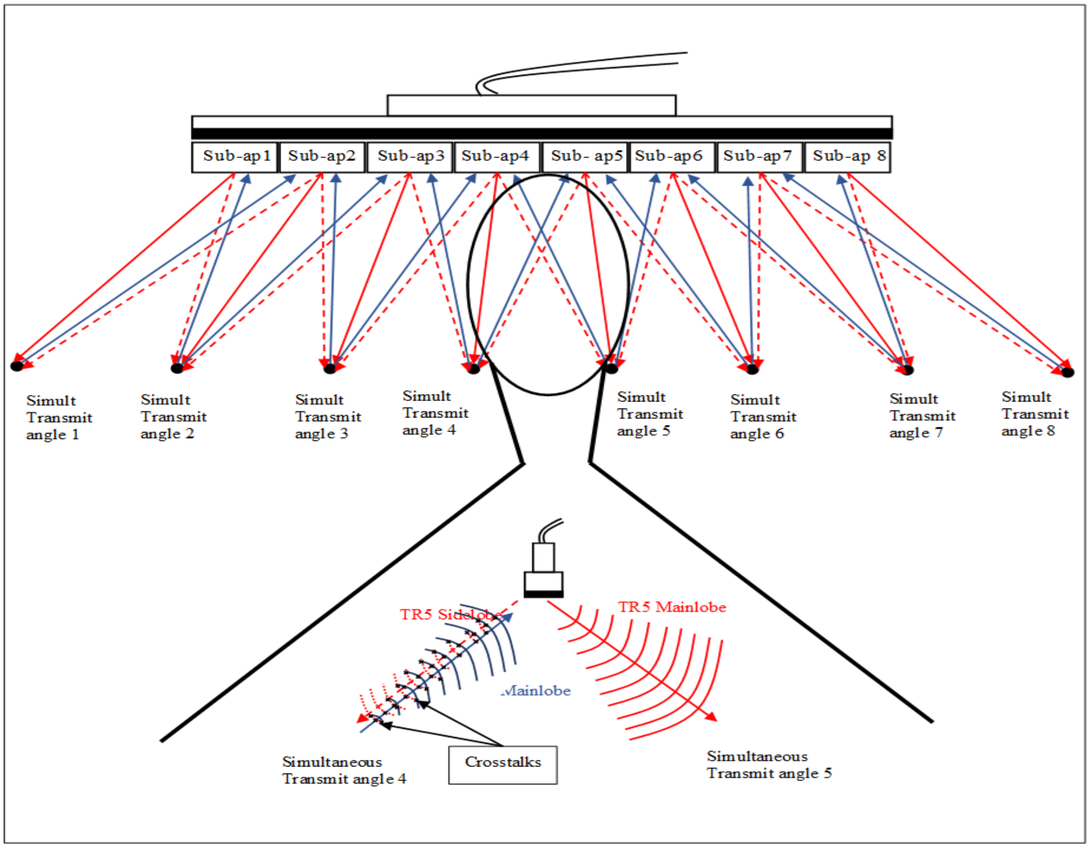


Fig. 1. Illustration of Transmission Crosstalk for same Focusing Depths.

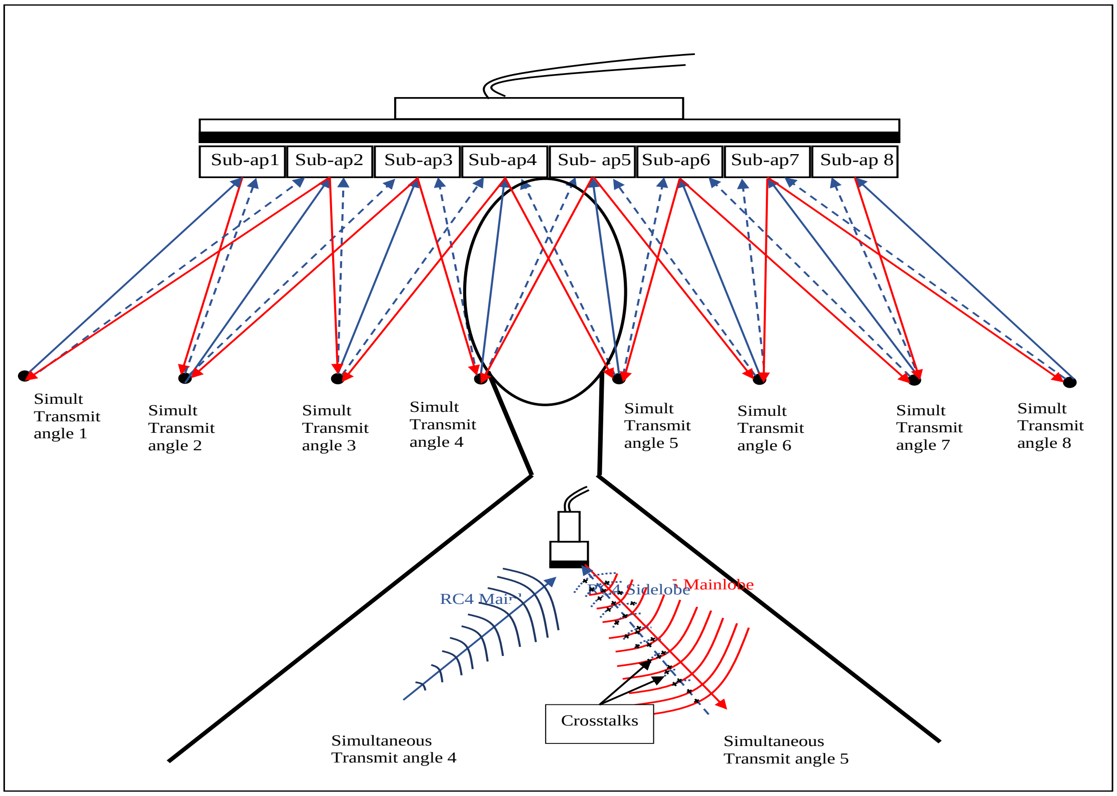


Fig. 2. Illustration of Reception Crosstalk for the same Focusing Depths.

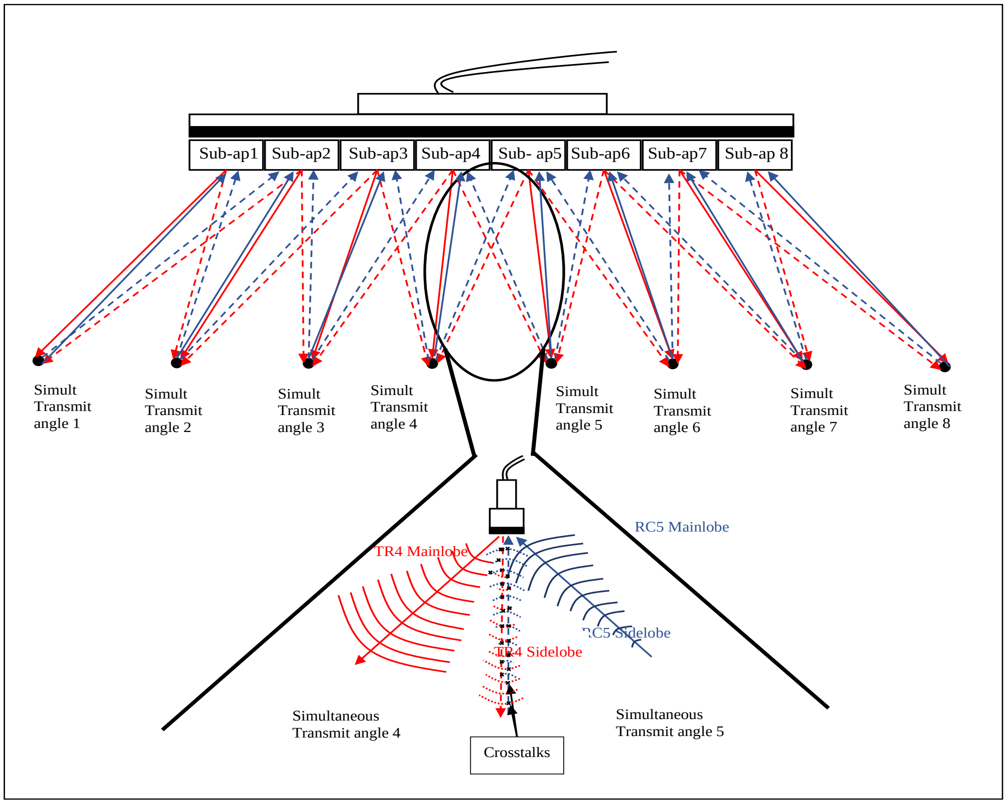


Fig. 3. Illustration of Transmission / Reception Crosstalk for the same Focusing Depths.

III. A New Approach for a Full MLT Sector Scan Geometry for 2D Echocardiography

A. Proposed Algorithm

In this work, we propose to benefit from the sector scan geometry to drastically reduce the crosstalk in the MLT technique. We talk about Synthetic Sum of full MLT (SS-MLT) low resolution images' algorithm. The four main ideas of our proposed algorithm are:

- Focusing at different depths in the same transmission event to considerably reduce main lobe crosstalk between two adjacent transmission lines. In other words, every transmission line focuses on the specific depth where the focusing points formed a diagonal line parallel to the diagonal direction of the sector scan image. Fig. 4 Illustrates Transmission Crosstalk reduction when each transmission sub-aperture focuses at different depths from other sub-apertures. We observe that the friction between the transmission side's lobe of TR5 and the reception of the main lobe of an adjacent RC4 is decreased. Also, the side lobe of the transmission direction is approaching to its original direction, so the distance between this one and the reception field of the adjacent focusing line is broadened. The restrained contact between the transmission side lobe and the reception of the main lobe results in a dramatic reduction of the transmission crosstalks compared to the traditional MLT algorithm.

Fig. 5 Shows Reception Crosstalk's reduction using this process, we notice that the contact between the transmission main lobe of TR5 and reception side lobe of adjacent RC4 line is reduced. Also, we note that the angles of emissions are broadened because the side lobe of the reception direction is approaching to its original one. Because of these two reasons, the reception crosstalk possibility in the system is highly limited compared to reception crosstalk possibility in the same focusing depth emission.

Fig. 6 presents the explanation of transmission/Reception Crosstalk reduction with simultaneous transmissions focusing at different depths. We remark that the contact between the transmission side lobe of one line and the reception side lobe of an adjacent line is reduced. Also, we note that the angle between the two sidelobes is broadened because each side lobe is getting close to its original direction, and the angle between the main transmission lobe and the main reception one of adjacent fields is widened. All this makes transmission/reception crosstalks at its minimum level.

- Using sparse emitting sub-apertures, to avoid sidelobe interference between adjacent, synchronous and transmitted waves relative to each transmitted beam. Each sub-aperture is apodized with Hanning apodization as it is presented in Fig. 7. Fig. 8 combines focuses at different depths and sparse emitting subapertures that have as consequence a very tiny transmission/reception crosstalks. Therefore, sparse emission leads to a more widened angle between adjacent fields. We observe that the main lobe reception and adjacent transmission side lobe are spaced enough so that there is almost no overlap between them, and the crosstalks are almost eliminated. Fig. 9 demonstrates the same result of a further reduction of reception crosstalk. Hence, the transmission main lobe and the adjacent reception side lobe are very spaced. Also, a very important reduction in transmission/reception crosstalks is observed in Fig. 10.

- Maintaining a high image power by focusing, several times, on different depths at the same zone, so that we disturb energy at all the sectors' scan geometry. Here, after each transmission event, we obtain one low-resolution image. After the summation of all the low-resolution images, we construct one high-resolution image which is the result of focusing on the morphological geometry (Fig. 11).
- Maintaining a high frame rate: For creating a cardiac imaging system, we must work at an increased frame rate specification. Analytically, one transmission/reception event takes about 200 us for a penetration depth of 15 cm and a speed of sound of 1540 m-s . Thereby, to maintain a 300 Hz frame rate, we are limited by 17 transmission-reception events. We divide the sector scan into 8 zones and insonifying each zone with sparse sub-apertures twice.

B. Scan Zones' Division

The most common causes of artifacts through MLT are the transmission and reception crosstalks; an immediate outcome of the communication between each main lobe line and the extra waves side lobes. We can easily observe that main-lobe/side- communications between adjacent transmitted lines can be reduced when we set each focus point of each transmission line at a different depth from the other adjacent focus points.

An accurate decision is to disturb the various focusing points along the line that is parallel to one of the diagonals of the scan geometry sector, as represented in Fig. 12. We obtain 8 scan zones. Different zones can have different numbers of focusing points. The intersection between scan lines and diagonals forms a sub-scan zone. In each sub-zone we have a focus point.

C. Insonification Table

We perform two insonifications to form one zone imaging. After the two transmission /reception events, we construct one low-resolution image. When performing 16 transmission/reception events, we obtain 8 low-resolution images. These images are summed together to put a high-resolution image.

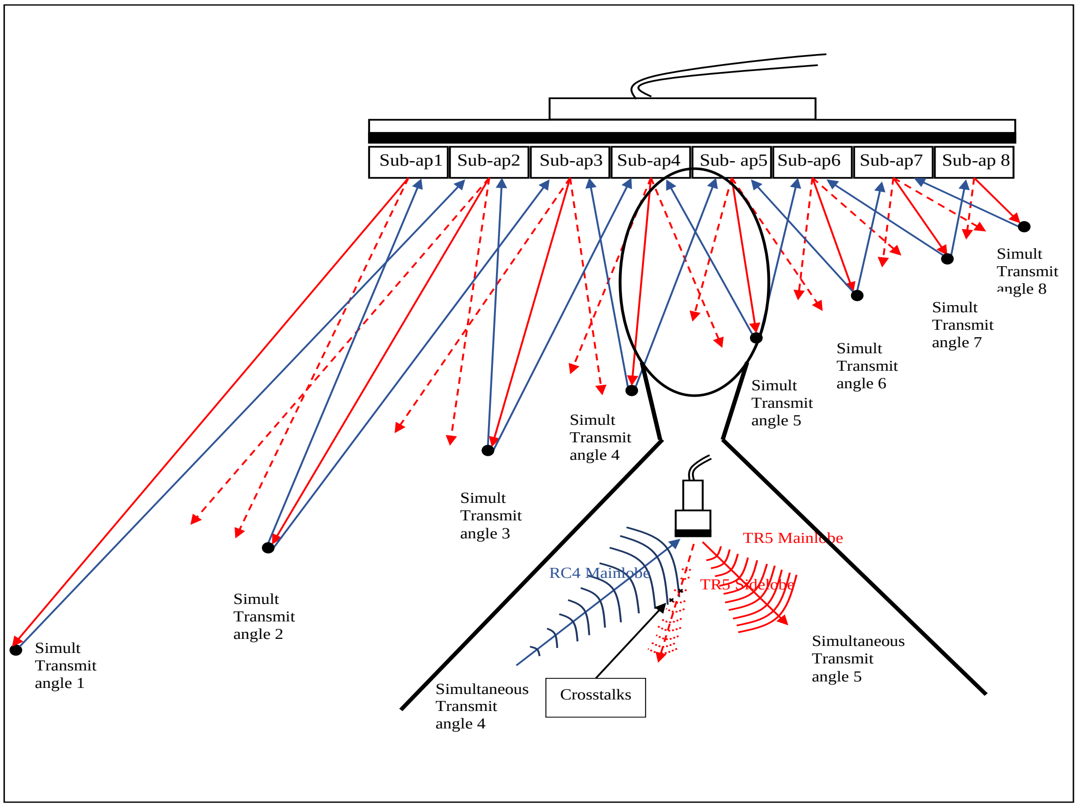


Fig. 4. Illustration of Transmission Crosstalk Reduction for different Focusing Depths.

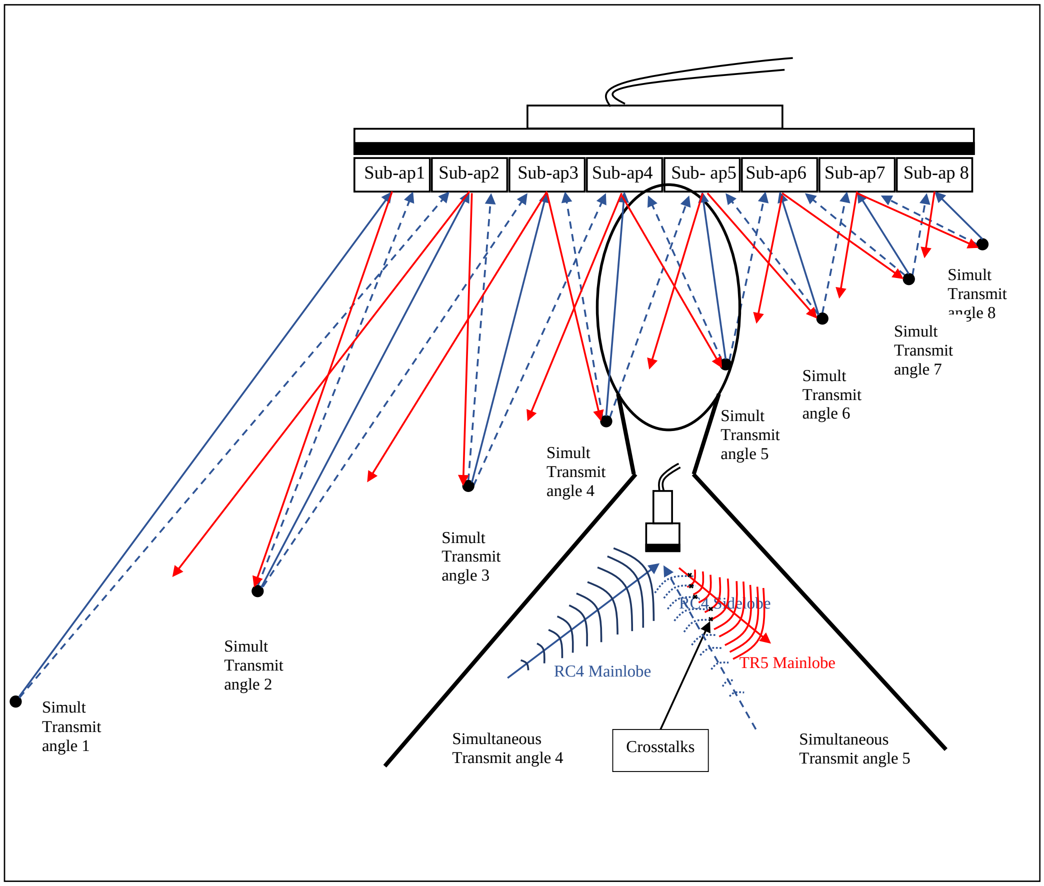


Fig. 5. Illustration of Reception Crosstalk Reduction for different Focusing Depths.

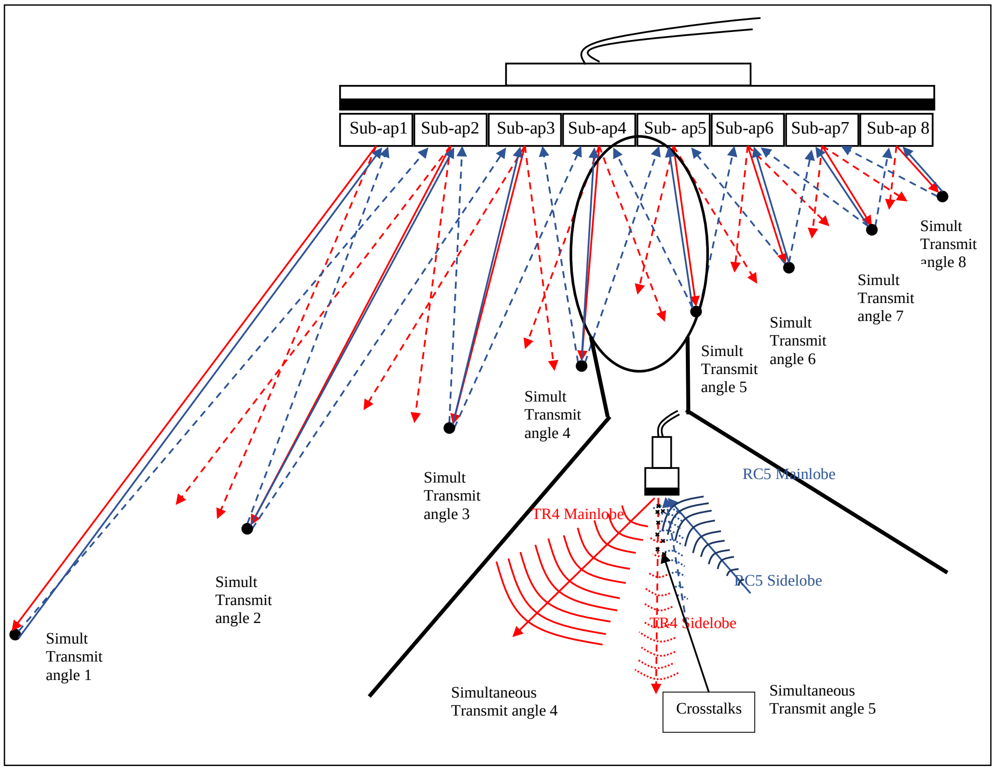


Fig. 6. Illustration of Transmission / Reception Crosstalk Reduction for different Focusing Depths.

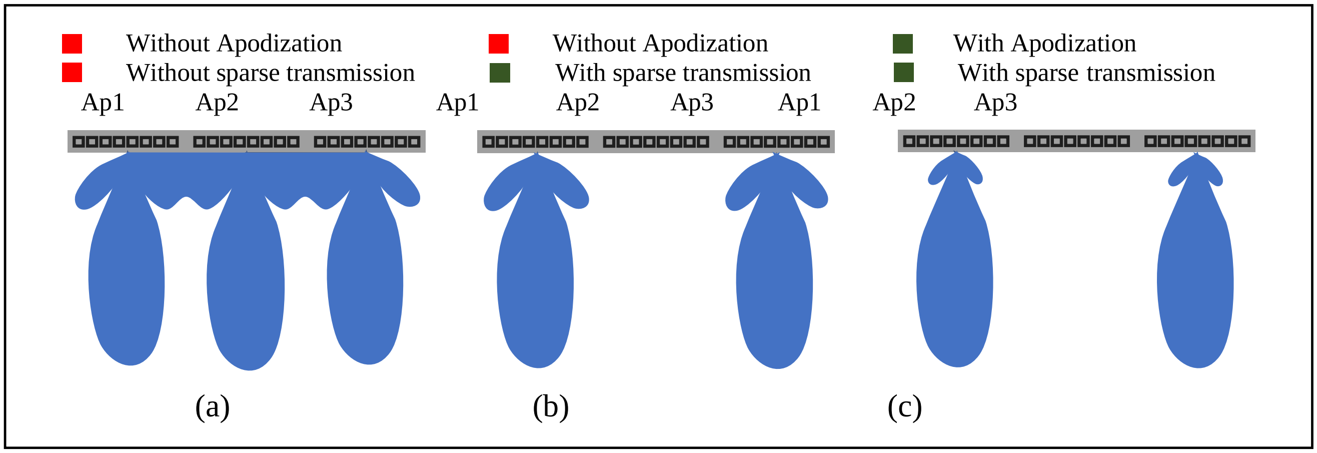


Fig. 7. Sidelobe Interference Reduction Steps (a) Sidelobe Interference Scheme before Apodization and without Sparse Transmission (b) Sidelobe Scheme before Apodization and after Sparse Transmission (c) Sidelobe Reduction Scheme after Apodization and after Sparse Transmission.

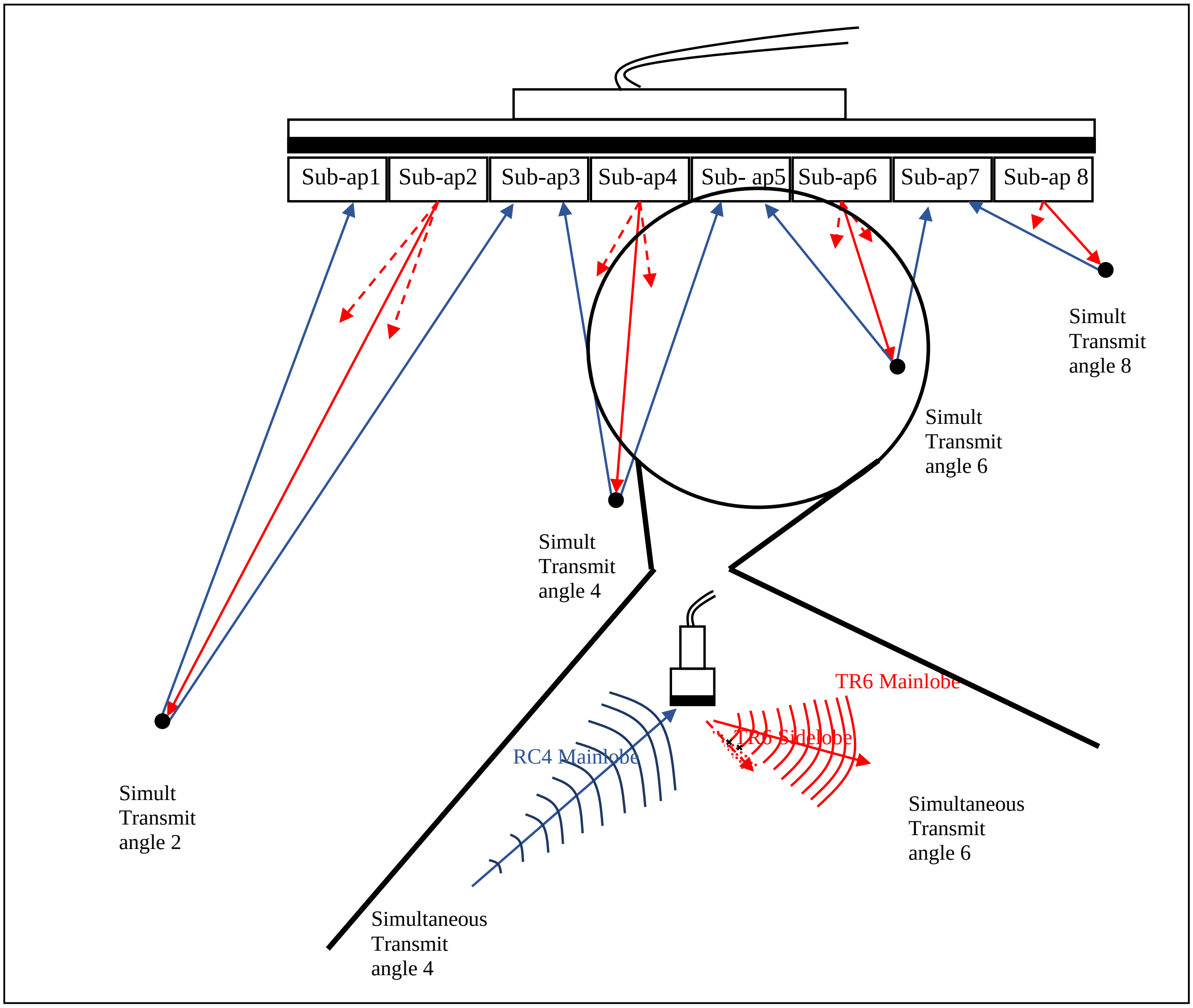


Fig. 8. Illustration of Transmission Crosstalk Reduction for different Focusing Depths and with Sparse Emitting Sub-apertures.

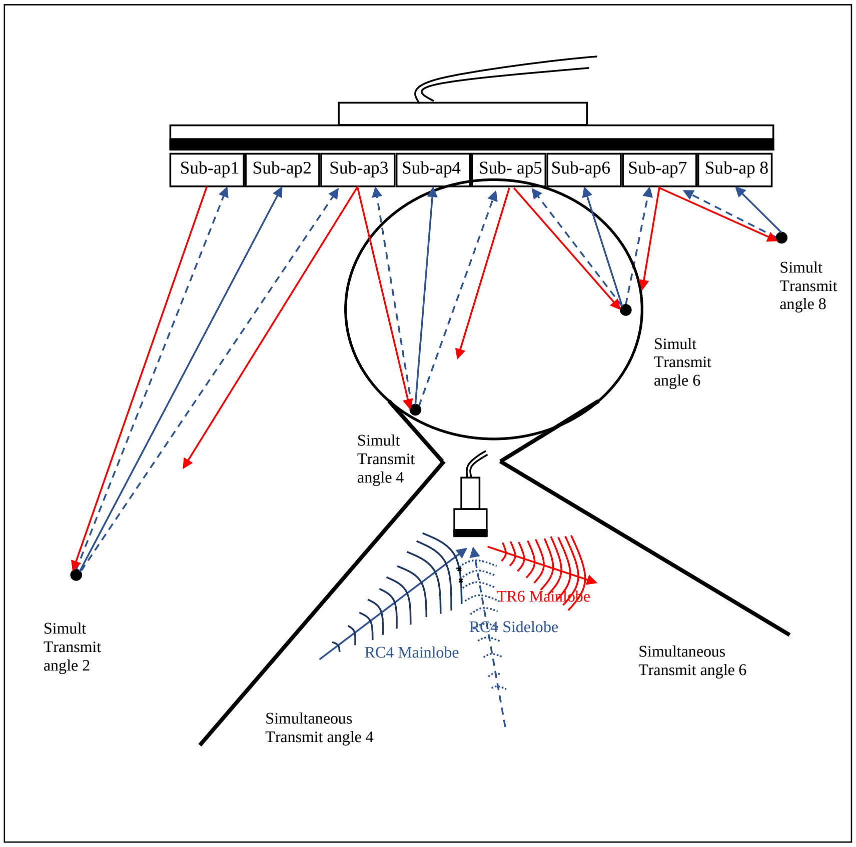


Fig. 9. Illustration of Reception Crosstalk Reduction for different Focusing Depths and with Sparse Emitting Sub-apertures.

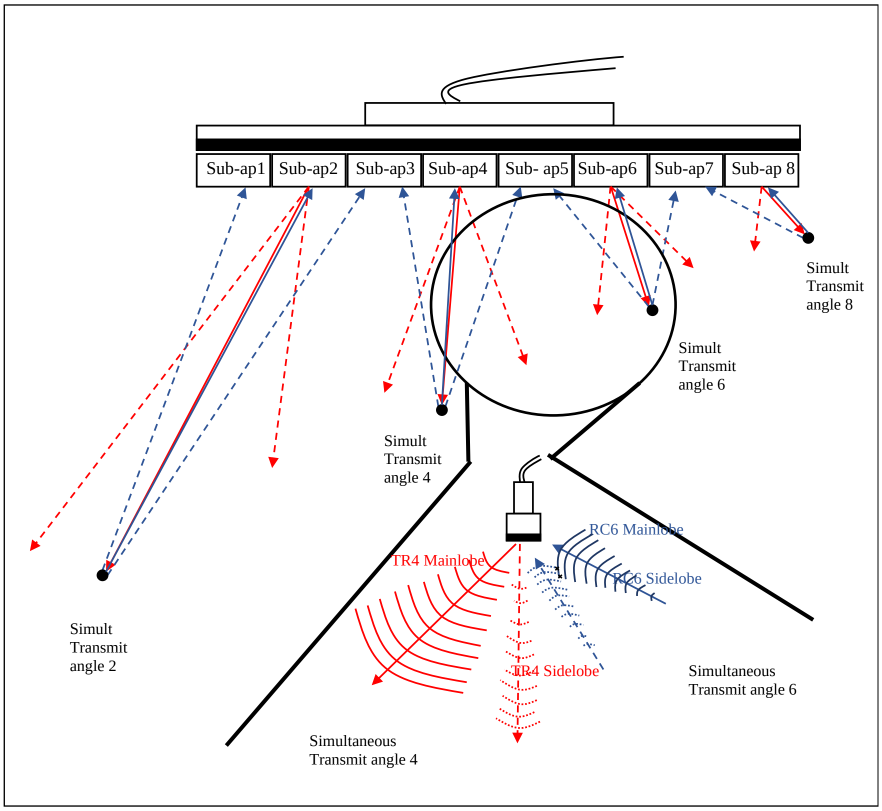


Fig. 10. Illustration of Transmission / Reception Crosstalk Reduction for different Focusing Depths and with Sparse Emitting Sub-apertures.

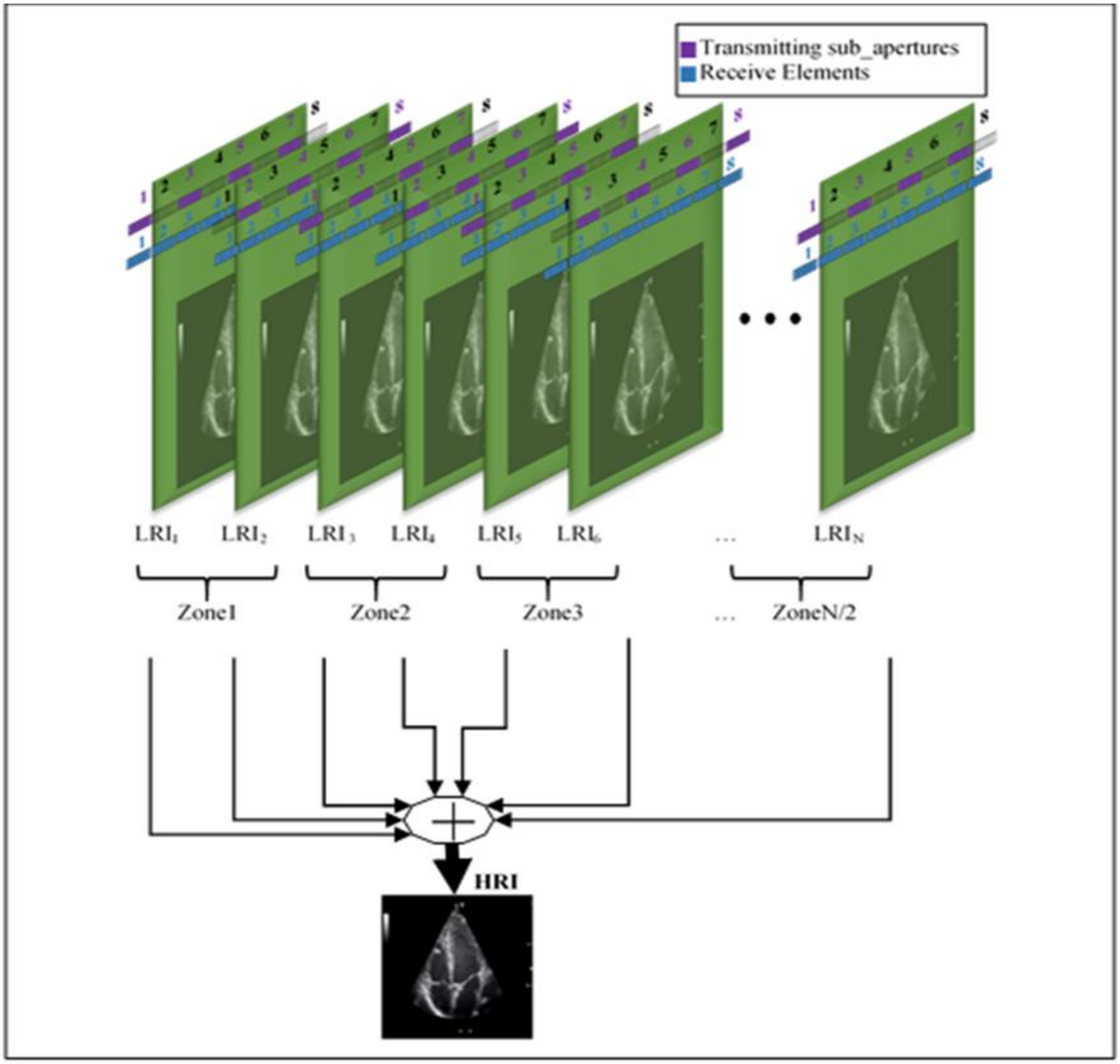


Fig. 11. High Resolution Image.

The insonification follows a certain firing order. For example, a focus point of Zone 1 and another in zone 5 are insonified at the same time. The same thing for a focus point of Zone 1 and another in zone 6 are insonified simultaneously.

Table I presents all the details about the focusing zones and the sub-apertures numbers and this is done for each emission number.

Here, zone 7 and zone 8 have one focusing point only. This decision is taken because the geometry surface spread over zones 7 and 8 is so small. So, we have decided here to gain frame rate. Also, we remark that (zone 2/zone 7+8), (zone 2/zone 6), (zone5/zone1) and (zone6/zone1) couples are highly spaced and echoes from these couples are temporally spaced. This makes the crosstalk highly reduced and maybe avoided between them. Here is the main idea of our new approach.

Another factor to talk about is the position of the focus centre for each transmission line per each transmission event. As you can deduce from the table below, the focusing point may be shared with multiple sub-apertures, here the position of the focusing point changes to be the centre of all the transmitting multi-elements' transducers. X1, X2,…X8 represent the fields that are perpendicular to the relative sub-apertures. For example, X1 is the field that is perpendicular to sub-aperture1, and so on to X8 which is perpendicular to sub-aperture 8. Fig. 13 and 14 describe the plotted emitted fields for each transmission synchronous line per each Low-resolution image.

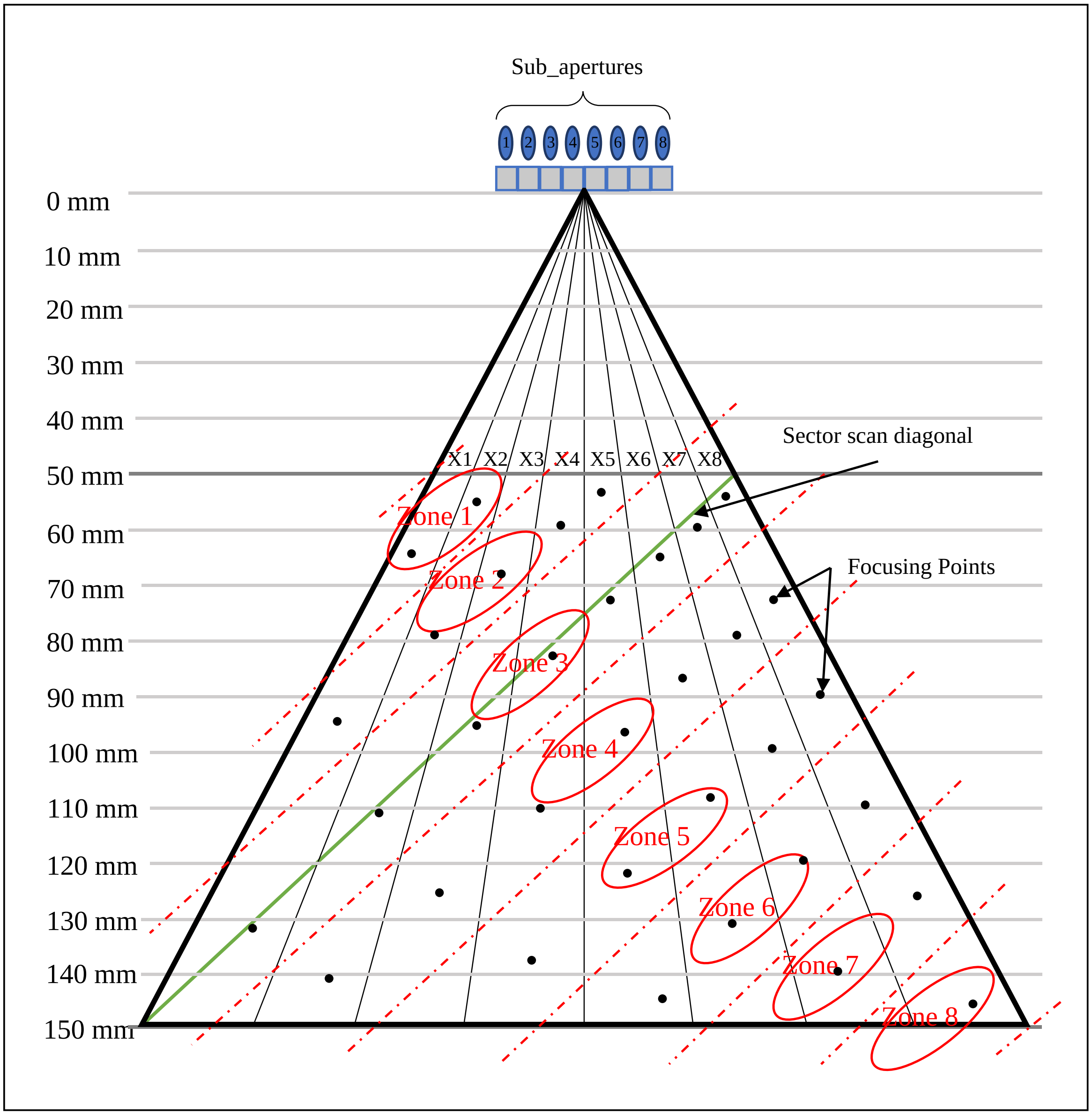


Fig. 12. Illustration of Synthetic Sum of Full MLT-MLA Scan Geometry (SSF-MLT-MLA-SG).

TABLE I. INSONIFICATIONS TABLE

Emission Number	**Focusing Zone Number**	**Sub_apertures (SbA)**							
		SbA 1	***SbA2***	***SbA 3***	***SbA 4***	***SbA 5***	***SbA 6***	***SbA7***	***SbA 8***
1	Zone2+7+8	95*X1	-	70*X3	-	55*X5	-	130*X8	
2	Zone 2 +6	80*X2		-	60*X4	-	130*X6	-	110*X8
3	Zone 3	130*X1	-	95*X3	-	75*X5	-	60*X7	
4	Zone 3	110*X2		-	85*X4	-	65*X6	-	55*X8
5	Zone 4	140*X2		-	110*X4	-	85*X6	-	75*X8
6	Zone 4	125*X3			-	95*X5	-	80*X7	
7	Zone 5	135*X4				-	110*X6	-	90*X8
8	Zone 5+1	55*X2		-	120*X5		-	100*X7	
9	Zone 6+1	65*X1		-	145*X5		-	120*X7	

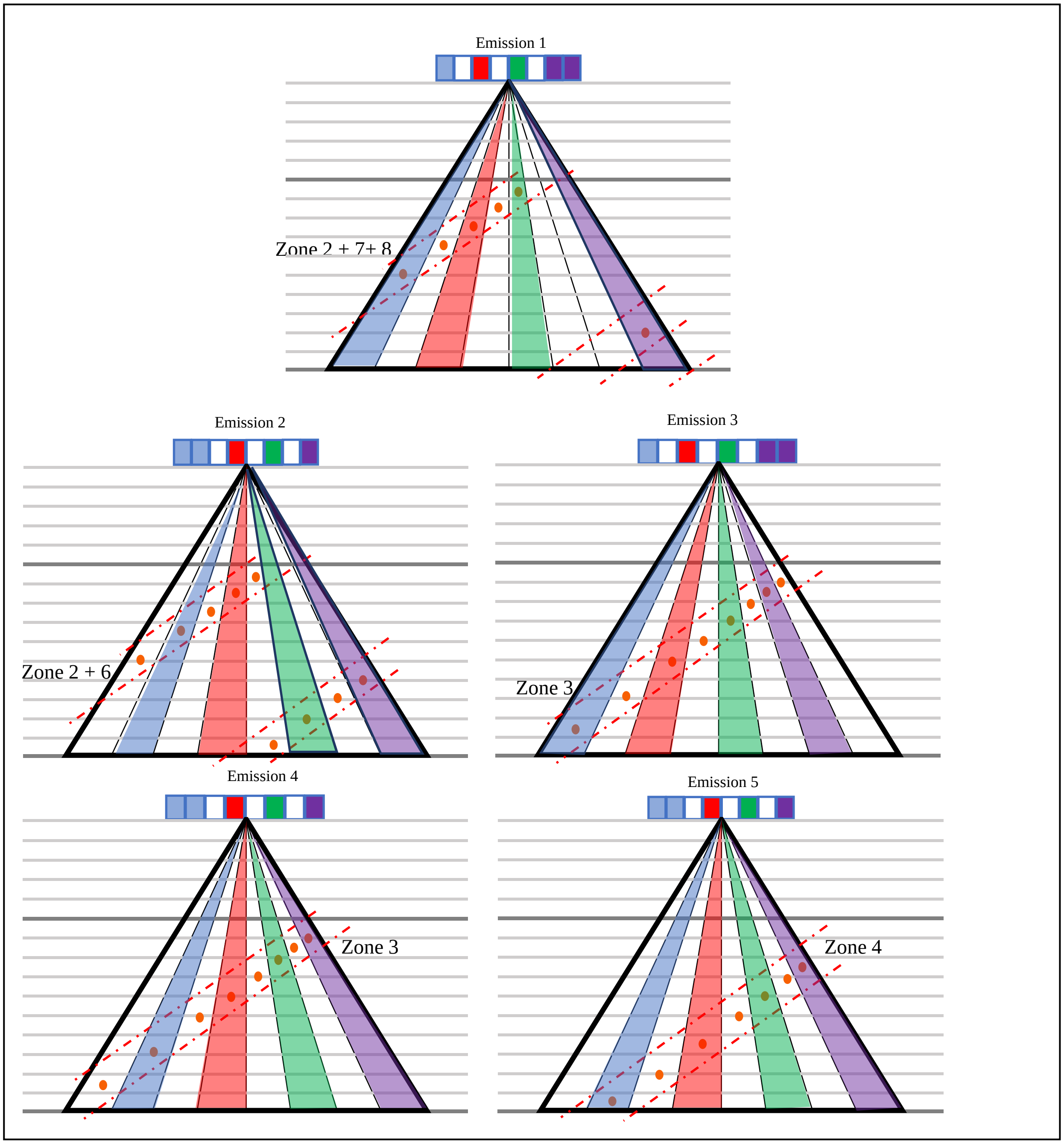


Fig. 13. Morphology of Focusing Point for each LRI: Emissions 1 to 5.

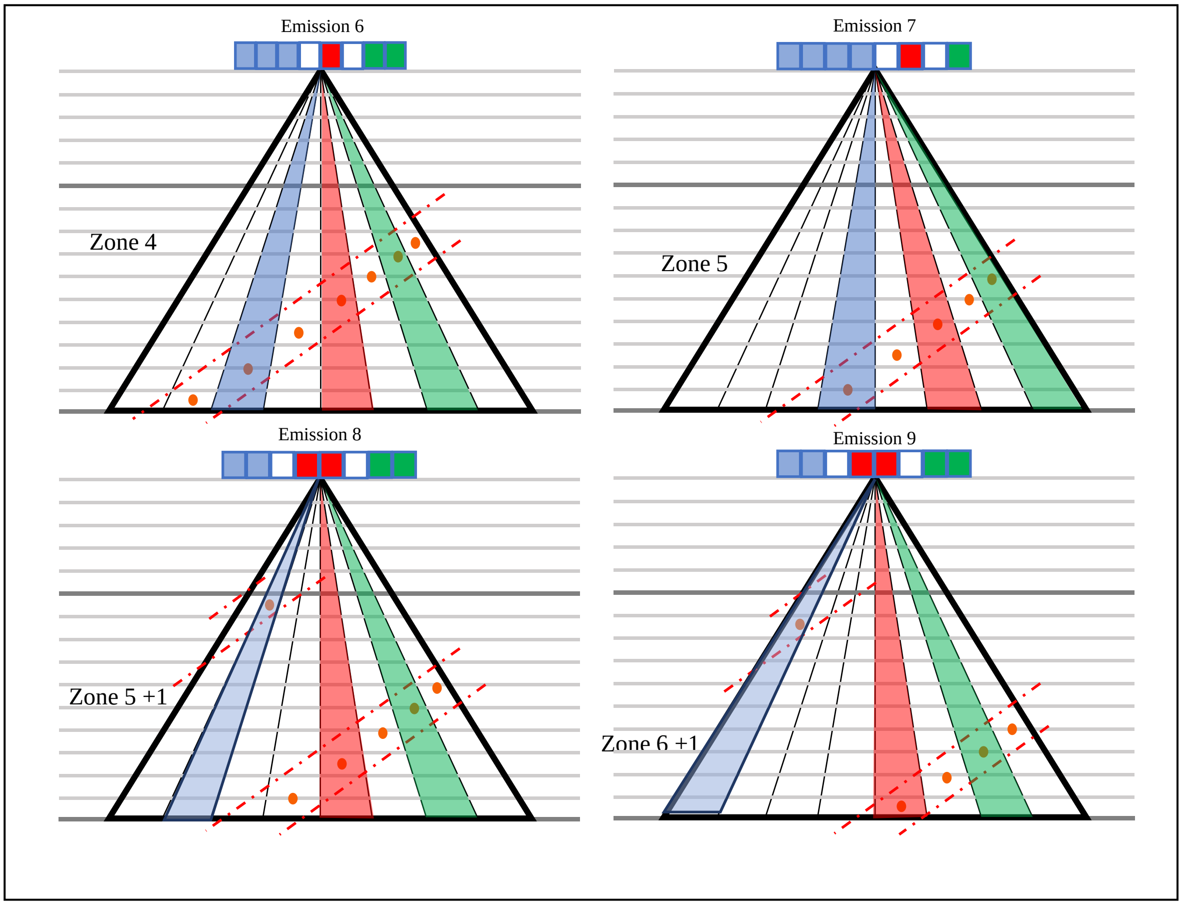


Fig. 14. Morphology of Focusing Point for each LRI: Emissions 6 to 9.

IV. Simulation Results

A. Simulation Setup

Each of the Point spread functions and the cyst phantoms are modelled. All the simulations are implemented in Matlab (MathWorks, Natick, Massachusetts) using Field_II. In the ultrasound software simulator [26], echo RF (Radio Frequency) data are produced using a 1D phased array transducer, which is composed of 128 elements with an element pitch of 193 µm and an element height of 5mm and an element width of 183 um and a kerf between adjacent elements of 10 um. This transducer is considered to scan a 30-image sectors using a speed of sound of 1540 m/s. Through transmission, the transducer generates a Hanning-windowed, 2-cycle sinusoid burst at 4 MHz and a relative bandwidth of 60% is used on transmission and the hanning apodization is applied accordingly. The same aperture with a hanning apodization is used in reception. The returned signals for each channel are sampled at 100 MHz. The RF data are imported to Matlab for beam forming. Dynamic reception focus is used for all the algorithms and the different beamforming techniques are applied to the recorded US signals. We use Hilbert transformation to the raw received RF data to detect the envelope and log compression. To prevent a coarse display due to a large image line spacing, we perform a low-pass interpolation to the raw received RF data laterally by a factor of 5. The scan conversion is in the last step to form the targetted image. All the images are represented over a 60-dB dynamic range.

B. Implemented Imaging Algorithms

To compare our proposed algorithm with the state-of-the-art 2-D cardiac imaging ones, subsequent beamforming methods are simulated. Their technical settings and detailed parameterization are described in the following paragraphs.

SLT in which the focusing points for each insonified beam are located at a penetration depth of 100 mm where all aperture elements are fired for each transmitting cycle. The focus points are equitably spread on a cylindrical shape centered at the middle of the transducer width, for a radius of 100 mm. The image lines are 128. Dynamic reception beamforming is exactly the mode used in this technique. Therefore, one transmission/reception event is employed to construct one scan line. Thus, for constructing one frame, one insonification is used line by line taking 26 milliseconds. So, the frame rate is approximately 40Hz (images per second).

8MLT beams are transmitted into 8 equidistant regions, each of them is splited into of 30/8 degrees opening angle. Focusing points are located at 100 mm depth for all parallel transmitting lines. After each insonification, 8 scan lines equal to the number of MLT beams are reconstructed simultaneously. To reconstruct an image of 128 lines, 16 firings are needed, and the resulting frame rate is 320frames per second (320 Hz).

16MLA beams are generated from 16 transmission events that focus on each time at the different regions, each of them is splitted into 30/8 degrees opening angle. Focusing points are located at 100 mm depth. After each insonification, 16 scan lines are reconstructed simultaneously. To generate an image of 128 lines, 8 firings are needed, and the resulting frame rate was 640frames per second (640 Hz). This imaging technique is simulated as a benchmark.

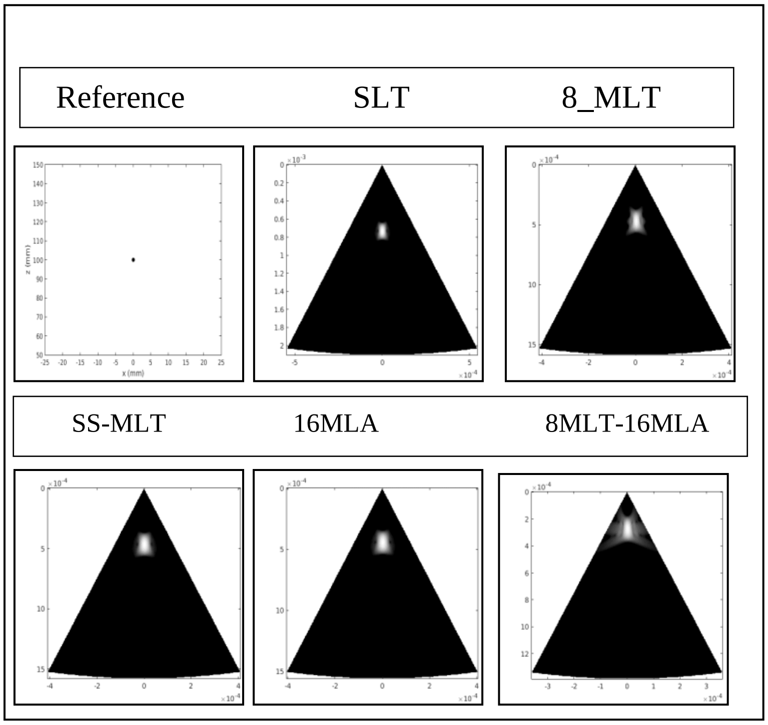
8MLT-16MLA beams are transmitted into 8 equidistant regions and 16 multilines are received. To reconstruct an image of 128 lines, 8MLT beams are transmitted into 8 equidistant regions, each of them is split into 30/8 degrees opening angle. Focusing points are located at 100 mm depth for all parallel transmitting lines. 16 MLA related to each region are reconstructed simultaneously around every transmission line, so 128 lines are generated for all MLT lines. Therefore, all scan lines forming one frame are generated simultaneously. Here we have a full image acquisition. One frame takes only one transmission/reception event, and the resulting frame rate is5128frames per second (5128 Hz).

In our algorithm, one High-resolution image is created after the Synthetic Sum of 9 low-resolution images. Every LRI is reconstructed with 128 MLA and a maximum with 4MLT beams that are transmitted into 4 equidistant azimuth planes with an inter-plane opening angle of 30/4 degrees. All MLT beams concerning each sub-aperture are focused at different depths, so, the first transmission line focuses on 70 mm and the second on 80 mm and so on and so forth to the last eighth line that focuses on 140 mm. The dynamic reception focus is adapted using the hole transducer elements to generate 128 MLA. For constructing one frame (HRI), we perform 9 transmission events equivalents to 9 LRI. The resulting frame rate is 5128/9frames per second (570 Hz).

C. PSF Results

The images of point spread are situated at 100 mm of depth obtained by SLT, 8_MLT, 16MLA, 8_MLT_16_MLAand SS-MLT beam formers are shown in Fig. 15, and the corresponding contour plots of the intensity point spread function is shown in Fig. 16. We can observe, in Fig. 15, that the SLT beam former exhibits the best performance and the best image quality output. We can notice in the same figure that 8_MLT beamformer shows approximately a performance that is a little bit better than our new proposed algorithm SS-MLT. After that, we remark that 16MLA occupies the fifth position concerning the image quality performance. Finally, we pay attention to 8_MLT_16_MLA algorithm that shows in the same figure also, the worst image quality. Fig. 16 is another form of displaying scan converted point spread function, so it exhibits the same analysis as the one described above.

Table II details the PSF performance for axial resolution. It shows the Full Width Half Maximum (FWHM), Full Width Half Dynamic Range (FWHDR), Main Lobe Width (MLW), Peak Side Lobe (PSL) and Double Main Lobe (DML) image quality options. The best parameters for each beamforming algorithm are highlighted in the table below. We notice that the best performances are obtained for the SS-MLT algorithm for which the values are respectively 58.6 for MLW, 82.6 for FWHM and 171.1 for FWDR. Similarly, SS-MLT has excellent values for PSL which is of the order of -503.609 as well as good values for DML which is of the order of -69.1 dB.


Reference SLT 8_MLT

SS-MLT 16MLA 8MLT-16MLA

Fig. 15. PSF Results.

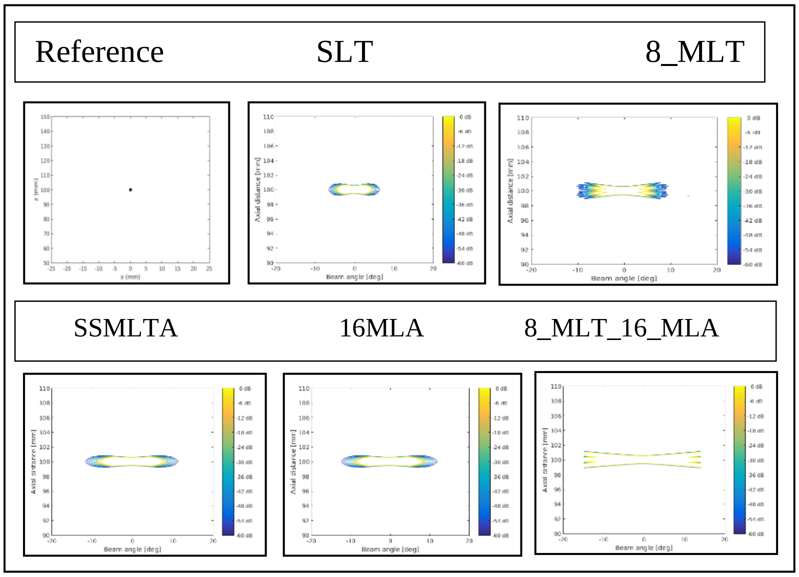


Fig. 16. PSF Contour Results

TABLE II. POINT SPREAD FUNCTION (AXIAL RESOLUTION)

Algorithm	**MLW (μm)**	**FWHM (μm)**	**FWHDR (μm)**	**PSL (dB)**	**DML (dB)**
SLT	**58.5**	**82**	**171**	**-550.511**	-68.6
8_MLT	**56.6**	**79.4**	**169.1**	-329.955	-67.9
SS-MLT	**58.6**	**82.6**	**171.1**	**-503.901**	**-69.1**
16MLA	58.6	82.8	171.3	**-498.884**	**-69.1**
8_MLT_16_MLA	59.3	83.3	171.9	-231.191	**-69.8**

In conclusion, SS-MLT has a very similar axial profile with the SLT, whose side lobes are almost always a little higher than those of the corresponding SLT, indicating a better image contrast resolution. The MLW for SS-MLT at a depth of 100 mm is less than that of the SLT, resulting in a wider FWHM and FWHDR than the SLT as shown in Table II.

Table III details the PSF performance metrics corresponding to the lateral resolution. Best MLW values are 57.6 µm, 44.4 µm, and, 66.8 µm for SLT, 8_MLT, and SS-MLT algorithms respectively. The best FWHM values are 82.8 µm, 55.3 µm, and 95.6 µm for SLT, 8_MLT, and SS-MLT algorithms respectively. Also, excellent FWHDR values are 184.5 µm, 97.1 µm, and 216.8 µm for SLT, 8_MLT, and SS-MLT algorithms respectively. The perfect PSL results are -378.469 dB, 0 dB, and -360.002 dB for SLT, 8_MLT, and SS-MLT algorithms respectively. The best DML results are -62 dB, -73.7 dB, and -60.2 dB for SLT, 8_MLT, and SS-MLT algorithms respectively.

This shows the same choices for image quality as axial resolution. Also illustrated are the optimal values of the three parameters for all techniques. SS-MLT showed a similar lateral profile with the SLT and 8-MLT, the PSL and DML of which are almost always marginally higher than that of the corresponding SLT and 8-MLT, suggesting an increased spatial resolution of the image contrast. The MLW for SS-MLT at a depth of 100 mm is a small greater than SLT and 8-MLT, resulting in a strong FWHM and FWHDR.

D. PSFs Results

The images of several point spread functions situated at different depths, in the [50mm-150 mm] range, obtained by SLT, 8_MLT, 16MLA, 8_MLT16_MLAand SS-MLT beamforming methods are shown in Fig. 17. We can remark that the similar image quality performance order like one point spread function could be noticed.

The efficiency metrics of PSFs compared to axial resolution are outlined in Table IV. For SLT, SS-MLT, and 8_MLT_16 MLA algorithms, the best MLW values are 59.6 mm, 58.9 mm, and 59.3 mm. The best FWHM values for SLT, SS-MLT, and 16MLA algorithms are 83.3 mm, 82.2 mm, and 83.2 mm respectively. Even for SLT, SS-MLT, and 16MLA algorithms, outstanding FWHDR values are 171.6 mm, 171.3 mm, and 172.2 mm respectively. For SS-MLT, 16MLA, and 8_MLT_16 MLA algorithms, the ideal PSL outcomes are -1,260 dB, -1,493 dB, and -1,412 dB, respectively. For SLT, SS-MLT, and 8 MLT 16 MLA algorithms, the best DML results are -69.8 dB, -69.1 dB, and -69.5 dB.

TABLE III. POINT SPREAD FUNCTION (LATERAL RESOLUTION)

Algorithm	*MLW (µm)*	*FWHM (µm)*	*FWHDR (µm)*	*PSL (dB)*	*DML (dB)*
SLT	**57.6**	**82.8**	**184.5**	**-378.469**	**-62**
8_MLT	**44.4**	**55.3**	**97.1**	**-7.3896e-13**	**-73.7**
SS-MLT	**66.8**	**95.6**	**216.8**	**-360.002**	**-60.2**
16MLA	84.1	120.3	272.2	-356.123	-59.7
8_MLT_16_MLA	94.9	134.9	305.3	-128.235	-59.8

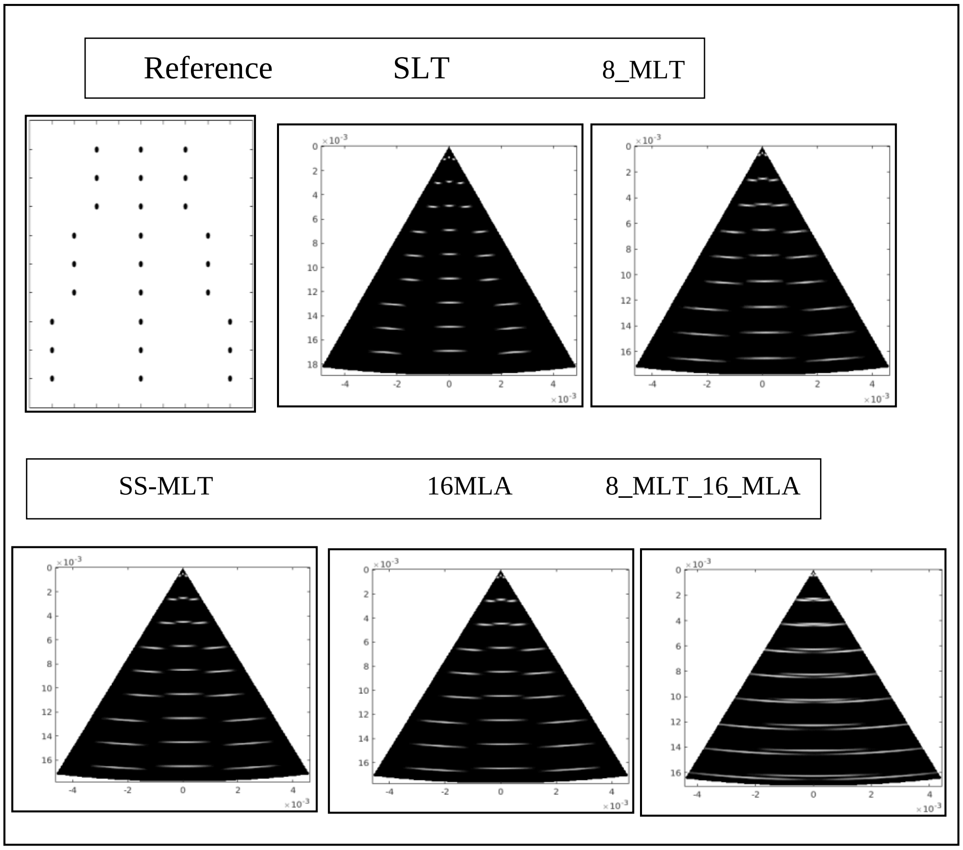


Fig. 17. PSFs Results.

TABLE IV. POINT SPREAD FUNCTIONS (AXIAL RESOLUTION)

Algorithm	*MLW (µm)*	*FWHM (µm)*	*FWHDR (µm)*	*PSL (dB)*	*DML (dB)*
SLT	**59.6**	**83.3**	**171.6**	-0.535	**-69.8**
8_MLT	59.8	84.3	174.6	-0.312	-68.8
SS-MLT	**58.9**	**83.2**	**171.3**	**-1.260**	**-69.1**
16MLA	59.7	**82.2**	**172.2**	**-1.493**	-68.5
8_MLT_16_MLA	**59.3**	83.5	172.7	**-1.412**	**-69.5**

Table V details the performance metrics of PSFs relative to lateral resolution. The best MLW values are 32.5 mm, 72.7 mm, and 59.9 mm for the 8-MLT, SS-MLT, and 8_MLT_16_MLA algorithms. The highest FWHM values are 49.5 mm, 102.9 mm, and 130 mm for the 8 MLT, SS-MLT and 16MLA algorithms, respectively. Excellent FWHDR values are 133 mm, 227.6 mm and 186.7 mm also for 8-MLT, SS-MLT and 16MLA algorithms. For 16MLA, SS-MLT, and 8_MLT_16_MLA algorithms, the optimal PSL outcomes are -1,260 dB, -1,493 dB, and -1,412 dB. For SLT, SS-MLT, and 8_MLT_16_MLA algorithms, the best DML results are -63.5 dB, -60.1 dB, and -60.4 dB, respectively.

TABLE V. POINT SPREAD FUNCTIONS (LATERAL RESOLUTION)

Algorithm	***MLW (µm)***	***FWHM (µm)***	***FWHDR (µm)***	***PSL (dB)***	***DML (dB)***
SLT	131	185.5	404.3	-0.535	**-63.5**
8_MLT	**32.5**	**49.5**	**133**	-0.312	-51.3
SS-MLT	**72.7**	**102.9**	**227.6**	**-1.493**	**-60.1**
16MLA	92.6	**130**	289.8	**-1.260**	-59.2
8_MLT_16_MLA	**59.9**	85	**186.7**	**-1.412**	**-60.4**

E. Cyst Phantom (CP) Results

Ten cysts with a radius that varies from 8.5mm to 2 mm are located at five depths of imaging to evaluate the beamformers under the cyst targets. The cyst phantom consists of a collection of point targets, five cyst regions, and five highly scattering regions. This can be used for characterizing the contrast-lesion detection capabilities of an imaging system. The scatterers in the phantom are generated by finding their random position within a 100 x 100 x 10 mm cube, and then ascribe a distributed Gaussian amplitude to each scatterer. If the scatterer resides within a cyst region, the amplitude is set to zero. Within the highly scattering region, the amplitude is multiplied by 20. The phantoms typically consist of 200,000 scatterers and simulating to 128 RF lines.

The reconstructed images are provided in Fig. 18. As it can be seen, the cyst targets are not well detected in the image generated by 8_MLT_16_MLA. Moreover, the reconstructed images are defected by the effects of the noise. Even though SS-MLT results in a higher-quality image and more detectable cyst targets, in comparison with 16MLA, the effects of the produced noise are still obvious. Besides, SLT suppresses the effects of the noise further, as was illustrated for wire targets, and results in a higher image quality in addition to8_MLT.

The PSNR, SNR and CPP parameters are determined for the reconstructed images for a quantitative analysis. The previous efficiency metrics for each cyst-using beamformer are listed in Table VI. For 8_MLT_16_MLA, 8 MLT, and SS-MLT algorithms, the best SNR values are 16.34 dB, 17.29 dB, and 16.90 dB. For 8_MLT_16_MLA, SS-MLT, and 16MLA algorithms, the optimal PSNR results are -29.91 dB, -30.46 dB, and -30.42 dB. We used SLT as a reference algorithm, to study the SNR and the PSNR for the other methods. The greatest CNR values are 54.24 dB, 54.70 dB, and, 54.15 dB for SLT, 8_MLT_16_MLA, and SS-MLT algorithms, respectively. Also, the best CPP values are 25.06 dB, 25.45 dB, and, 25.09 dB for SLT, 8-MLT, and 16MLA algorithms, respectively.

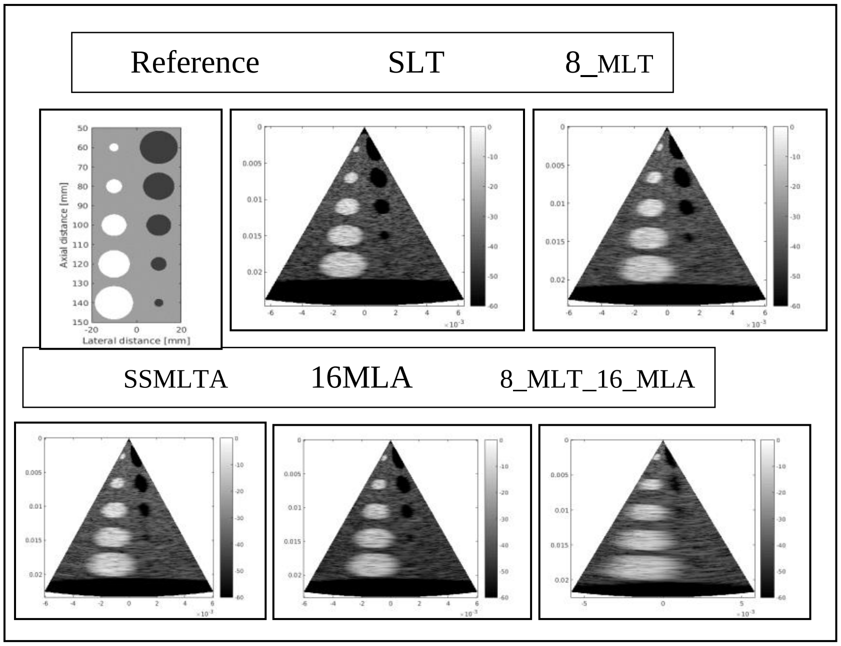


Fig. 18. Cyst_Phantom Results (200000 Scatterers).

TABLE VI. PERFORMANCE METRICS OF CYST PHANTOM

Algorithms	***SNR (dB)***	***PSNR (dB)***	***CNR (dB)***	***CPP***
SLT	Reference*	Reference*	**54.24**	**25.06**
8_MLT	**17.29**	-29.40	52.83	**25.45**
SS-MLT	**16.90**	**-30.46**	**54.15**	25.54
16MLA	16.22	**-30.42**	53.77	**25.09**
8_MLT_16_MLA	**16.34**	**-29.91**	**54.70**	25.74

*:SLT algorithm is taking as reference in comparison with other algorithms.

V. DISCUSSION

In this study, we brought to the fore a new implementation technique for a 2-D MLT system. As the MLT methods suffer from the crosstalk artifacts, we introduced many developments and enhancements in order to reduce the negative impact of these problems. Synthetic Sum of full MultiLine Transmission (SS-MLT) low-resolution images focused at diagonal focusing depths is our new proposed modality which allows for an improved image quality in 2D echocardiography imaging applications. Taking advantage of the diagonal synchronous focusing points, we highly decrease the interference among the different transmission/reception beams of the images and therefore reduce the amplitude of the crosstalk artifacts. Using sparse transmitting sub-apertures, we avoided transmission side-lobe/side-lobe beams interference between adjacent transmitted beams. Also, windowing functions adopted with the sub-aperture elements for both transmission and events might reduce the main-lobe/side-lobe interactions and equivalently reduce the transmission/reception crosstalk levels. In this work, we used Hanning apodization on transmission and on reception, which gives further perfections and improvements in the decreasing of the crosstalk artifacts for all simulations.

Simulation results prove that our new proposed method is effective and promising regarding point spread function performances, thus it competes for SLT image quality options for many configurations. Besides, for cyst phantom simulations, it has shown good CNR and SNR and this is done to the low crosstalk levels as we explained above. Table VII gives an approximative order of the algorithms referring to their operating frame rate and their image quality performances. Although the SS-MLT runs at a higher frame rate than 8MLT, it exhibits better competences than those of B-mode images.

TABLE VII. FRAME RATE AND IMAGE QUALITY BEST ORDER

Best Image Quality per Order	***Algorithm***	***Frame Rate (Hz)***
1	SLT	5128/128
3	8_MLT	5128/16
2	SS-MLT	5128/9
4	16MLA	5128/8
5	8MLT-16MLA	5128

VI. CONCLUSION

This paper demonstrates a new MLT-MLA technique that benefits from the shape of the sector scan geometry. We suggest distributing multiple focusing points along the diagonal of the sector scan geometry. The synthetic sum of multiple low-resolution images is adopted to bring a higher resolution image. Each LRI is constructed with MLT lines. By using sparse sub_apertures, the interference is reduced and by using different focusing depths the spatial resolution of the images is ameliorated with a good SNR and CNR and perfect FWHM lateral resolution and axial resolution values. Adequately, the crosstalk noise is drastically reduced while maintaining a high frame rate that allows 2D echocardiography imaging. After we have accurately studied all the conditions of implementation of our method, and after we concluded that it was very promising for cardiac applications, we concentrate now on its hardware implementation to further evaluate its reliability and effectiveness.

REFERENCES

[1] T. Szabo, Diagnostic Ultrasound Imaging—Inside Out. 2004.

[2] M. Cikes, L. Tong, G. R. Sutherland, and J. D’hooge, “Ultrafast cardiac ultrasound imaging: technical principles, applications, and clinical benefits,” JACC: Cardiovascular Imaging, vol. 7, pp. 812-823, 2014.

[3] H. Chen, T. Varghese, P. S. Rahko, and J. A. Zagzebski, “Ultrasound frame rate requirements for cardiac elastography: Experimental and in vivo results,” Ultrasonics, vol. 49, no. 1, pp. 98–111, Jan. 2009.

[4] H. Hasegawa, H. Kanai., “High-frame-rate echocardiography using diverging transmit beams and parallel receive beamforming,” J. Med. Ultrason, Vol. 38, pp. 129–140, Jul. 2001.

[5] L. Tong et al., “Comparison of Conventional Parallel Beamforming With Plane Wave and Diverging Wave Imaging for Cardiac Applications: A Simulation Study,” IEEE Trans. Ultrason. Ferr. Freq. Control, Vol. 59, pp. 1654–1663. 2012.

[6] Ø. Ragnhild. “Coherent Plane-Wave Compounding in Medical Ultrasound Imaging: Quality Investigation of 2D B-mode Images of Stationary and Moving Objects,” 2012.

[7] T. Shirasaka, “Ultrasonic imaging apparatus,” US Patent 4815043, Mar. 1989.

[8] L. Tong, H. Gao, and J. D’hooge “Multi-transmit beam forming for fast cardiac imaging: quantitative analysis of the cross-talk between MLT beams,” IEEE International Ultrasonics Symposium, Dresden, Germany, pp. 1271–1274, 2012.

[9] L. Tong, A. Ramalli, R. Jasaityte, P. Tortoli, and J. D’hooge, “Multitransmit beam forming for fast cardiac imaging–experimental validation and in vivo application,” IEEE Trans. Med. Imag., vol. 33, no. 6, pp. 1205–1219, Jun. 2014.

[10] L. Demi et al., “Implementation of Parallel Transmit Beamforming Using Orthogonal Division Multiplexing—Achievable Resolution and Interbeam Interference,” IEEE Trans. Ultrason. Ferroelectr. Freq. Control., Vol. 60, pp. 2310–2320, 2013.

[11] L. Demi et al., “Tissue harmonic images obtained with parallel transmit beamforming by means of orthogonal frequency division multiplexing,” In Proceedings of the 2014 IEEE International Ultrasonics Symposium (IUS)., 3-6 September 2014, Chicago, Illinois pp. 1213-1216, Sep. 2014.

[12] B. Denarie, H. Torp, T. Bjastad, “A Novel Approach for Reducing Multi Line Transmission Cross-talks using 2D Transducer Arrays,” IEEE International Ultrasonics Symposium, Dresden, Germany, pp. 2242 – 2245, 2012.

[13] B. Denarie, T. Bjastad, and H. Torp, “Multi-line transmission in 3-D with reduced crosstalk artifacts: A proof of concept study,” IEEE Trans. Ultrason. Ferroelectr. Freq. Control, vol. 60, no. 8, pp. 1708–1718, 2013.

[14] A. Ramalli, L. Tong, J. Luo, J. D’hooge, P. Tortoli, “Safety of fast cardiac imaging using multiple transmit beams: experimental verfication,” UltrasonicsSymp. (IUS), 2014 IEEE International, 2014.

[15] A. Ramalli et al., “Real-Time High-Frame-Rate Cardiac B-Mode and Tissue Doppler Imaging Based on Multiline Transmission and Multiline Acquisition,” IEEE Trans. Ultrason. Ferroelectr. Freq. Control, vol. 65, no. 11, pp. 2030–2041, Nov. 2018.

[16] P. Santos, L. Tong, A. Ortega, EigilSamset, and J. D’hooge, “Safety of Multi-line transmit beam forming for fast cardiac imaging – a simulation study,” UltrasonicsSymp. (IUS), 2014 IEEE International, 2014.

[17] L. Tong et al., “Fast three-dimensional ultrasound cardiac imaging using multi-transmit beamforming: A simulation study,” IEEE Ultrason. Symp. Proc., Vol. 60, pp. 1456–1459, 2013.

[18] L. Tong, H. Gao, and J. D’hooge, “Multi-transmit beam forming for fast cardiac imaging-a simulation study,” IEEE Trans. Ultrason. Ferroelectr. Freq. Control, vol. 60, no. 8, pp. 1719– 1731, Aug. 2013.

[19] L. Demi, M. D. Verweij, and K. W. A. Van Dongen, “Parallel transmit beamforming using orthogonal frequency division multiplexing applied to harmonic Imaging-A feasibility study,” IEEE Trans. Ultrason. Ferroelectr. Freq. Control, vol. 59, no. 11, pp. 2439–47, Nov. 2012.

[20] D. Dubberstein and O. V. Ramm, “Methods and systems for ultrasound scanning using spatially and spectrally separated transmit ultrasound beams,” U.S. Patent 6159153, Dec. 12, 2000.

[21] K. Thiele, “Multi-beam transmit isolation,” U.S. Patent 20100016725, Jan 21, 2010.

[22] J. Hossack and T. Sumanaweera, “Method and apparatus for medical diagnostic ultrasound real-time 3-D transmitting and imaging,” U.S. Patent 6179780, Jan. 30, 2001.

[23] C. Yoon, et al.,“Orthogonal quadratic chirp signals for simultaneous multi-zone focusing in medical ultrasound imaging.” IEEE Trans. Ultrason. Ferroelectr. Freq. Control, vol 59, pp. 1061–1069, 2012.

[24] F. Prieur, B. Dénarié, A. Austeng, and H. Torp, “Correspondence—Multi-line transmission in medical imaging using the second-harmonic signal,” IEEE Trans. Ultrason., Ferroelectr., Freq. Control, vol. 60, no. 12, pp. 2682–2692, Dec. 2013.

[25] L. Tong et al., “Coded excitation for crosstalk suppression in multi-line transmit beamforming: Simulation study and experimental validation,” Appl. Sci., vol. 9, no. 3, p. 486, Jan. 2019.

[26] J.A. Jensen., “Field: a program for simulating ultrasound systems,”10th Nordicbaltic conference on biomedical imaging., pp. 351-353, 1996.

Secure Energy Efficient Attack Resilient Routing Technique for Zone based Wireless Sensor Network

Venkateswara Rao M[1]
Research Scholar
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundaion
Vaddeswaram,Andhra Pradesh, India

Srinivas Malladi[2]
Professor
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram,Andhra Pradesh, India

Abstract—**Security and Energy efficiency are two key factors to be contemplated in the design of applications based on wireless sensor networks (WSNs). Optimization of energy consumption is obligatory for an increased life time of the network. Without security, attackers can disrupt the entire operation of sensor network by instigating diverse attacks like message tampering, message dropping either in partial or whole and message flooding etc. This work proposes a secure energy efficient attack resilient routing technique for zone based wireless sensor network with proactive detection of malicious zones and mitigation from the attacks. Different from earlier works on detecting each malicious node, this work cleaves the network to zones and allots a probabilistic fuzzy score to model the success ratio of packet propagation through the zone. The routing is adaptive to ongoing residual energy and security risks. A Firm decision cannot be made on the frameworks influencing the life time of the network considering it may influence the operations of network. Experimentation of the proposed solution is done in NS2 and contrasted with the existing solutions to prove the effectiveness of the approach.**

Keywords—Energy efficiency; malicious zones; preference score probabilistic fuzzy score; residual energy

I. Introduction

Wireless sensor network is a network of sensor associated with a wireless infrastructure. The sensing nodes senses assorted parameters and propagate it via multi hop to descend in the networks. With expeditious reduction in cost of sensors, WSN is used in numerous applications like precision farming, industrial security, wild life monitoring, etc. Typical WSN consists of nodes connected with wireless infrastructure. Most sensor network deployments are unattended with sensor nodes being powered with batteries.

Due to sensing, listening, processing, transmission and reception of packets, the battery energy [1] is consumed and these processes are not scheduled in optimum manner, the energy depletes at faster rate and node becomes dead. The node failure does not only reduce the sensing coverage area, but also affects the routing. The communication holes created in the network disrupts the multi hop routing and minimizes the network reliability. Due to unattended nature, it is not possible for battery replacement for dead nodes. In this situation prudent use of energy is the best possible way to increase the life time of nodes and the network. Due to Wireless infrastructure, WSNs are prone to various attacks like message drops, message tampering, denial of service through message flooding etc. Detection of these attackers [25] and mitigation from these attacks is essential for reliability of applications using the sensor network. Implementing a higher complexity security algorithm involves higher energy consumption [11] for processing at sensor nodes and this in-turn reduces the life time of the node.

This work proposes an adaptive security enforcement solution with energy efficiency in routing for wireless sensor networks. The routing protocol is designed to be resilient against message drop, message forging and message flooding attacks. The entire network is split to multiple zones and each zone is allocated to two scores i.e. security score and energy score. The security score which models the security available in that zone. The energy score is calculated based on initial energy of the zone. Preference score is calculated using security and energy scores. Based on the preference score the routing is adapted in such way to meet both security and energy consumption requirements of the applications. Fuzzy logic is applied to score the zones probabilistically based on the current packet delivery performances.

Further, the paper is illustrated as follows: The related work is discussed in Section II. Problem definition is elaborated in Section III. The proposed work is described in Section IV. Novelty of proposed work is presented in Section V. Results of the proposed work is depicted in Section VI. Conclusion and future enhancements are discussed in Section VII.

II. Related Work

A routing protocol using inter cluster coordination with a goal of energy efficiency is developed in [1]. Received signal strength (RSS) from base station is used for clustering decision. A selection protocol to choose the cluster coordinators from node to sink is proposed to transmit data. in [2]. Energy consumption is reduced with efficient clustering with Gaussian Elimination algorithm with the goal to increase network life time. QoS based routing protocol for body area networks is proposed in [3]. Delay is calculated for all paths from source to sink and best path satisfying the delay requirements of the application is selected for routing. A multi path optimized routing protocol is proposed in [4]. Multi base stations to reduce route hops and ON-OFF cycles are the two strategies adopted to increase life time. Energy aware routing protocol is proposed in [5]. It is specifically designed for query-based applications. Zonal broadcasting is adopted in this work to reduce the total energy consumption. QOS routing based on multiple constraints in proposed in [6]. The protocol is highly scalable and adaptive to network dynamics. Greedy forwarder selection approach based on energy aware aggregated metric is proposed to select the best hop for routing. Network overhead is reduced in this protocol due to reduction in exchange of control packets. A scalable routing protocol with the goal of increasing network life time is proposed in [7] and [13]. Analytical model is proposed to find the optimum number of clusters based on location of base station and distribution of nodes. Best clusters to route the packets are selected based on current residual energy. Optimal relay node for packet forwarding based on multiple constraints is proposed in [8] to select the efficient relay nodes for packet forwarding. Packets are differentiated based on the priority. Differential path selection for different packet priority is done based on delay and current residual energy. Energy efficient routing is realized using Artificial Neural Network (ANN) in [9]. Next hop is selected using ANN. QOS factors and residual energy is used to train the ANN. Threshold sensitive Energy Efficient sensor Network protocol (TEEN) is proposed in [10]. The protocol works best for time critical applications. The sensing rate is controlled by the applications based on available energy in the network. Frequent data need requirements are satisfied using this protocol. Routing protocol using the concept of directional antenna is proposed in [12] with the goal for achieving energy efficiency. It is built on top of DSR protocol with extensions for power efficient gathering. Genetic algorithm is combined with bacterial foraging optimization to select energy efficient paths for data transmission. But sensors power consumption due to use of the hybrid algorithm is not considered in this work. An approach for reducing the packet transmissions and the energy consumption due to using the spatial and temporal correlation in the data is proposed in [14]. Temporal correlations in the sensor data is found using prediction-based approaches. Data transmission frequency is reduced for sensors with high correlations in their data. Energy efficient multi path routing protocol is proposed in [15]. Multiple paths are found using particle swarm optimization (PSO). The parameters to find efficient path is optimized using neural network. But the network overhead is very high in this approach due to frequent flooding. A heuristic solution is proposed in [16] to detect grayhole attackers. AODV is extended with bait detection schemes to find grayhole attackers. Since attackers are detected during route discovery stage, the network overhead for detection during packet forwarding is avoided. Black holes are detected using cooperative sensing in [17]. Due to the detection of Black holes, there is limitation in network life time and packet delivery ratio. PSO based clustering is proposed in [18]. Cluster heads were selected based on multiple criteria like mobility, energy. The life time of the network improved using fitness function adopted in this approach. Geographic routing protocol is combined with PSO with the goal of increasing life time of sensor networks in [19]. Energy utilization is drastically reduced in this approach. The fitness function for PSO considers inter node distance and nodal power. Security against routing attacks is studied in [20]. The work proposed an algorithm to detect cooperative black hole attacks. The routing protocol is controlled by some designated nodes in the network called security monitoring nodes (SMN). The black holes in the network are detected by the SMN and routing path is prevented from falling to black hole nodes. However, the work did not consider SMN compromise and there is a limitation in average energy of nodes. Packet transmission delay is also more in this works PSO is combined with TORA routing protocol in [21]. The problem of energy efficient route selection is treated as an optimization problems and PSO algorithm searches for the optimal solution for the problem. Load is distributed fairly and life time of network is increased using this solution. In clustering based solutions, the energy of nodes near to sink is depleted fast. This problem is referred as hotspot. Authors in [22] proposed sink mobility-based solution to solve the hotspot problem. Leech clustering protocol is optimized for energy efficiency in [23]. To overcome the demerits of predominant methods and further increase the lifetime of WSNs, a novel improved energy-efficient LEACH (IEE-LEACH) routing protocol is schemed in [23]. Malicious nodes are detected in the network using combination of cooperative bait detection and reverse tracing in [24]. Neighbor based Cluster Location Aware Routing (NCLAR) is modeled to achieve more packet delivery rate with high location accuracy in [25].Implement a framework for detection of malicious nodes in WSN and elimination of Node in actor nodes by creating a topological structure dynamical by adapting a connectivity point of peer to peer and point to point communications is proposed in [26]. A clustering and localization techniques for improving an efficient energy routing method are schemed in [27]. Certificate revocation based malicious attacker detection and prevention is proposed in [28]. Authors in [29] proposed weight-based clustering with a goal of increasing the life time of the network. An optimized trust-based ant colony optimization (TACO) and integrity verification techniques were implemented for wireless node initialization and trust probability computation in [30]. The author in [31] describes the minimum energy broadcast (MEB) problem for increasing the life time of nodes in wireless sensor networks. Secure communication model was proposed in [32] by considering time, energy, and traffic as factors for wireless sensor networks. The clustering of multiple paths in wireless sensor network is depicted in [33] by combinedly finding the membership and the number of clusters using SCAMS (simultaneous clustering and model selection). An efficient Multiple Objective Function (OF) in Routing Protocol for Low Power and Loss Network (RPL) is introduced in [34] for a Wireless Sensor Network in the field of healthcare. The important factors to be considered for OF are Packet Delivery Ratio (PDR) and the power consumption of sensors.

III. Problem Description

Given a wireless sensor network with N nodes and single sink, each node has residual energy of E and there are k attackers distributed across the network. The attacker can be message dropper (black hole and gray hole attack), message tamper and message flooder (Denial of service attack). Based on the current residual energy E, the sensor nodes are configured with the data sensing rate. Multi hop routing is used to forward data from nodes to sink. First node death concept is used to measure the life time of the sensor network. The objective of this work is to have a secured energy efficient route from source to sink resilient against the attacks.

IV. Secure Energy Efficient Attack Resilient Routing

Architecture is given in the Fig. 2 for proposed solution. Every sensor node contains a unique ID and it is preconfigured with Hyperelliptic curve cryptography (HCC) private key and the corresponding public key is kept at sink node. The key pair between the node and sink is unique and it is not available to any other nodes in WSN.

Hyperelliptic curves are generalization of elliptic curves. In these curves, the value of genus is greater than 1. ECC is special case of Hyperelliptic curve with genus value as 1. Compared to ECC, not much work has been done on Hyperelliptic curves. The work so far done in Hyperelliptic curves are only academic domain.

Assume k be a field. The general equation of hyperelliptic curve C with genus g over k

$C{:}y^2{+}a(x)y{=}b(x)$

Where

a(x) is a polynomial of degree $\leq$ g over b

b(x) is a monic polynomial of degree 2g+1 over b

As an example, following is a sample HCC function

$C{:}y^2 = x^5 - 5x^3 - 4x - 1$ over Q genus g=2

The number of nonintersecting simple closed curves that can be drawn on the surface without separating it is known as Genus of a curve. The number of handles is equal to the genus of a curve. The corresponding non intersecting closed curves are shown in Fig. 1.

Due to offline generation and pre-configuration of keys, there is no energy consumption due to processing of key generation and sharing. Every node is also assigned with a secret key sequence and a hash function H which is known only between the node and the sink.

The entire network is divided into $M \times M$ zones. Every zone has zone ID. The zone id is given by the sink for every zone as a number increasing from 1 to M*M.

The size of the zone is set in such a way that nodes in the same zone are within one hop with other nodes in that zone. All the nodes in the zone operate in duty cycle, so that each node in its duty cycle can process the packet for that zone and forward it to next hop. The duty cycle for the nodes is preconfigured after deployment.

$Zone\ size = \ comm.range^2$

Sink keeps the mapping of nodes in the zone. For each zone, sink maintains two scores – Security Score, Energy Score. Both scores are in value of 0 to 10. Security score of 10 means, there are no known attacks in the zone and score of 0 means there is a severe security risk for the zone. Energy score of 10 means that there is a higher residual energy in the zone and score of 0 means the zone is dead. In addition to these scores, sink maintains following counters for each zone

1) Packet traversal count (PTC)
2) Packet traversal failed count (PTFC)
3) Tampering incident count (TIC)
4) No Tampering count (NTC)
5) Total packets passed (TPC)

All these counters are set to 0 initially.

The AODV routing protocol is extended to design the proposed protocol. The RREQ and RREP message is added with additional fields for the extended routing protocol.

RREP and RREQ modifications are represented in Table I and Table II, respectively.

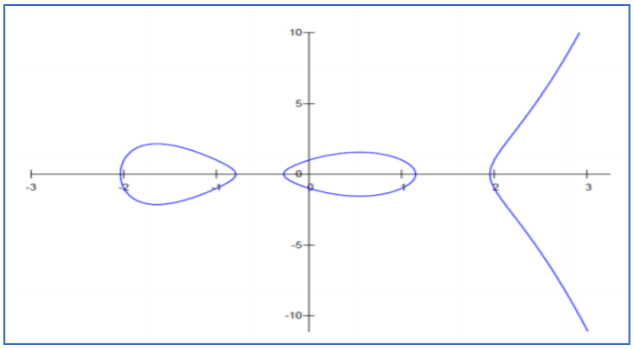


Fig. 1. HCC Curve.

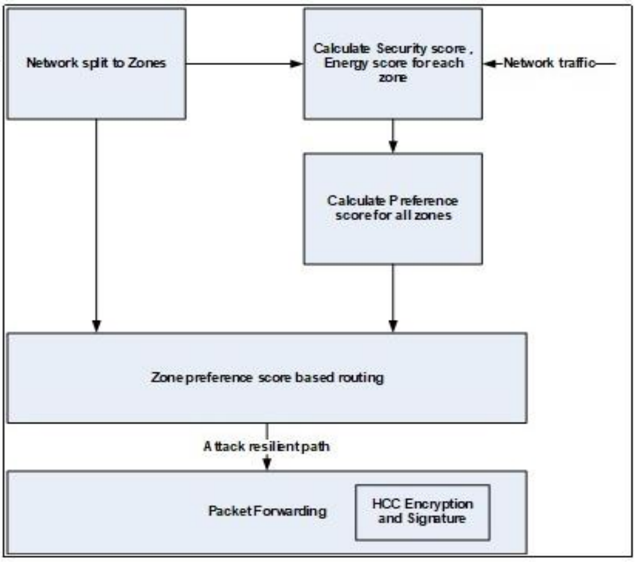

Network split to Zones
Calculate Security score , Energy score for each zone
Network traffic
Calculate Preference score for all zones
Zone preference score based routing
Attack resilient path
Packet Forwarding
HCC Encryption and Signature

Fig. 2. Architecture.

TABLE I. RREQ MODIFICATIONS

Field Name	Detail
Reqkey	Hashed key is added in this field

TABLE II. RREP MODIFICATIONS

Field Name	Detail
Encrypted PS score	Preference score is encrypted and stored
Signature	This signature is used for verification the RREP at the source

When a node wants to route packet, it instigates a RREQ request. The last sent secret key sequence is hashed using H and this hashed sequence is added to the RREQ request. RREQ request reaches the sink through multiple paths. At sink, analysis of each of path is done and a preference score is calculated for each route.

The analysis of path in the RREQ request at the sink node involves following steps:

1) Each node in the path is mapped to the corresponding zone.

2) The security score of the path is the average of the security score of all zones in the path.

3) The energy score of the path is minimum of the energy scores of all zones in the path.

The security score of the path is calculated as

$$SS_P = \frac{\sum_{i=1}^{N} SS_i}{N}$$

Where SS_i is the security score of the zones of N zones in the path.

The energy score of the path is calculated as

$$ES_p = \prod_{\min of\ all\ N} ES_i$$

Where ES_i is the energy score of the path.

A preference score is calculated for the path as weighted function of security score and energy score.

$$PS = w_1 * SS_p + w_2 * ES_p$$

With $w_1 + w_2 = 1$

Sink applies the hash function H on the secret key sequence in the RREQ and uses it to encrypt the preference score using AES encryption function. Also hash of the encrypted preference score and path in the RREP and the key sequence is done using H and this is inserted as signature in the RREP reply and sent. Sink maintains the paths and the preference score generated for a RREQ request from the source in its memory for a certain time period T. We refer this as cache.

Once the source node receives the RREP, it verifies the signature to check if the RREP message was tempered by modifying the path and the preference score. The signature verification is done by generating a signature as below and checking whether the generated signature is similar to the signature in the RREP message.

$$LS = H(\ last\ sent\ secret\ key\ sequence)$$

$$Sig_g = H(encrypted\ PS\ in\ RREP\ |path\ in\ RREP\ |\ LS)$$

$$\begin{cases} Sig_g == Sig\ in\ RREP\ , RREP\ is\ valid \\ Sig_g \neq Sig\ in\ RREP\ , RREP\ is\ invalid \end{cases}$$

The source nodes take the path with highest preference score among valid RREP and use it for routing the packets.

When sink receives the data packets, it finds the path taken for routing from the data packet. If the path is same as the highest preference score path in the Cache, then it increments the PTC count for all zones corresponding to the nodes in the path. If the path is not same, it means that path did not reach the source or the message would have got tampered. In this case, the PTFC count is incremented for all zones corresponding to nodes in that highest preference path and PTC count in incremented for all the zones corresponding to nodes through which the data packet has arrived. All data packets sent from source to sink is signed with the same procedure followed for RREP. At sink the digital signature is verified and if the verification fails, the TIC count is incremented and if the verification succeeds, NTC count is incremented. These counters are increased only once for the data session and not for all the data packets in the session. But TPC count is incremented for all zones corresponding to the nodes in the path every time data packet is received in the sink.

Sink calculates the average rate of data session from a zone and the deviation from the average rate is decided as flooding. The source node from whom data session exceeded the threshold is decided as flooding attacker. To prevent other nodes energy getting exhausted due to processing of flooded packet, sink sends black list packet with the information of flooding node to all the neighbor zones of the flooded node. The neighboring zones drops the packets from black listed flooders. By this way, flooding attack is mitigated in the proposed solution.

The energy score for the zone is calculated based on the TPC count as follows:

$$ES = \frac{10 * (E - TPC * \ E_c)}{E}$$

Where the initial energy of the node is E and the energy consumed for transmission and reception of a packet at node is E_c.

The security score is modelled as fuzzy function of following inputs:

1) Packet traversal count (PTC)
2) Packet traversal failed count (PTFC)
3) Tampering incident count (TIC)
4) No Tampering count (NTC)

These four input variables are fuzzified into three ranges using transfer function. These ranges are represented as Low(L), Medium(M) and High(H) using transform function shown in the following Fig. 3, 4, 5 and 6 for PTC, PTFC, TIC, NTC respectively.

The output variable of security score is also expressed in the form of transform function for values from 0 to 10 in terms of three variables of Low (L), Medium (M) and High (H). The transfer function for the output security score is given in Fig. 7.

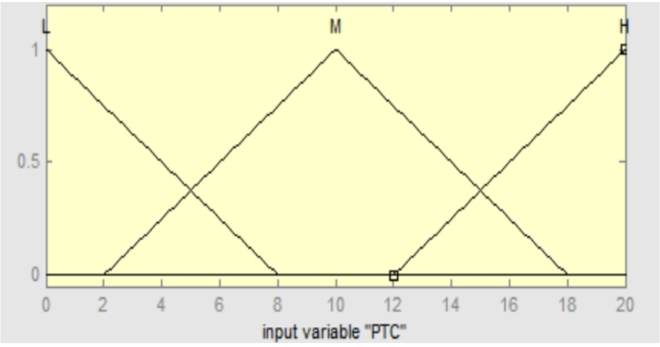


Fig. 3. Fuzzified Transfer Function for PTC.

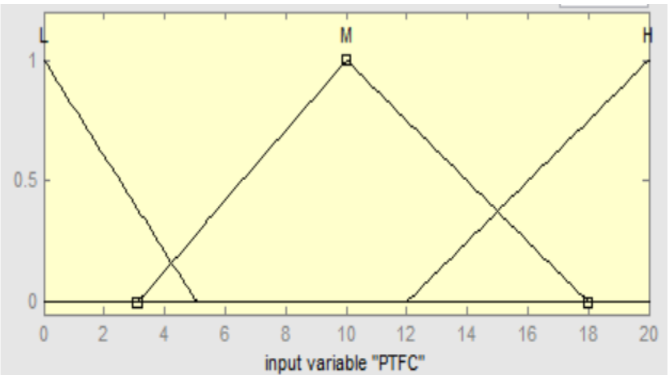


Fig. 4. Fuzzified Transfer Function for PTFC.

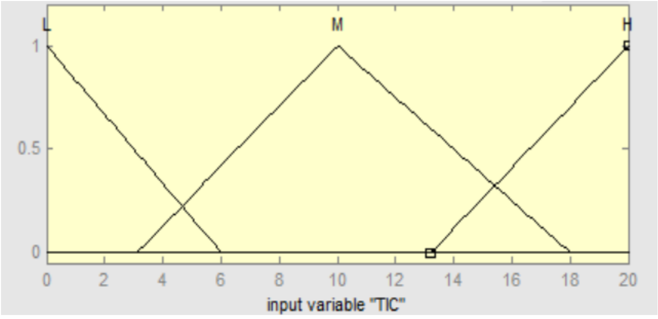


Fig. 5. Fuzzified Transfer Function for TIC.

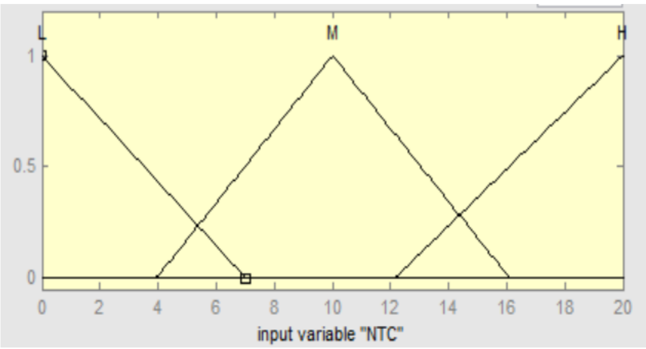


Fig. 6. Fuzzified Transfer Function for NTC.

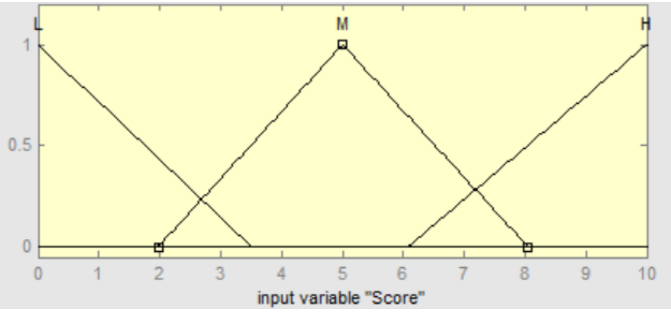


Fig. 7. Fuzzified Transfer Function for Security Score.

Fuzzy rule is designed by mapping the input variables to the output variable for all combinations of inputs. With this fuzzy system a probability fuzzy score for security is calculated from 0 to 10 using the four inputs of PTC, PTFC, TIC and NTC.

The security score calculated using fuzzy function is given as

$$F(score) = \mu_1 * Q(PTC) + \mu_2 * Q(PTFC) + \mu_3 * Q(TIC) + \mu_4 * Q(NTC)$$

Here The fuzzification kernel for input x is Q(x).

The center of gravity method is applied to get defuzzification.

$$Score = \frac{\int \mu^{-}_{Dr}(x).\, xdx}{\int \mu^{-}_{Dr}(x).\, dx}$$

Where x = {PTC, PTFC, TIC, NTC}

Compared to other attack detection schemes, in the proposed scheme, we are not detecting the individual attacker. But with presence of attackers in the zone, the security score value goes very low and this zone becomes an unpreferred choice in routing. By this way the routing protocol is resilient against attacks.

The flow chart for secure energy efficient attack resilient routing is shown in Fig. 8.

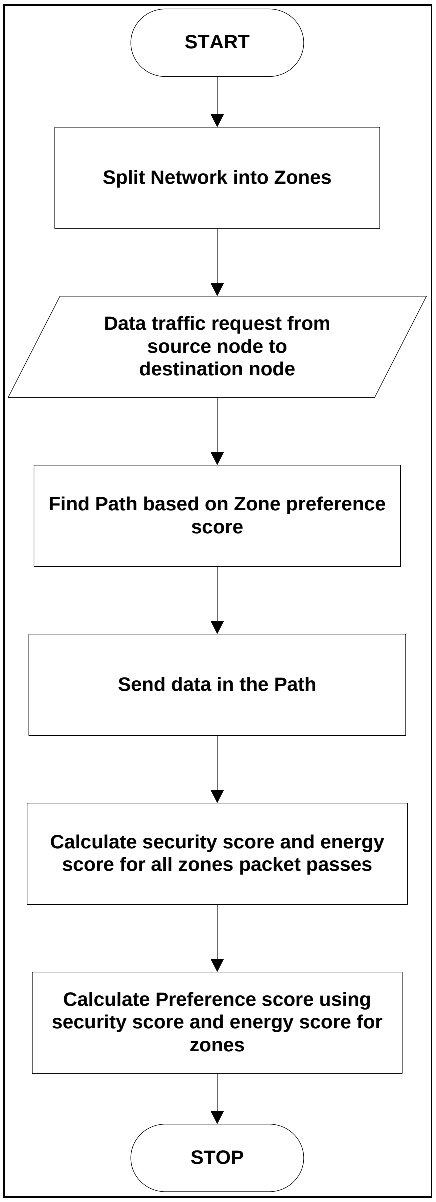

START
Split Network into Zones
Data traffic request from source node to destination node
Find Path based on Zone preference score
Send data in the Path
Calculate security score and energy score for all zones packet passes
Calculate Preference score using security score and energy score for zones
STOP

Fig. 8. Flowchart.

V. Novelty in Proposed Solution

The proposed solution is different from existing solutions is following aspects:

1) Instead of detecting individual attackers, it models the security in terms of zones and moves all the complexity to sink instead of nodes.

2) Paths are adaptively scored based on the security risk and the current residual energy.

3) Multiple attacks of message drop, forging and flooding is considered.

4) Tampering of RREP message contents is also considered.

5) Packet delivery ratio and Life time of the network is improved.

6) Average energy of nodes is increased.

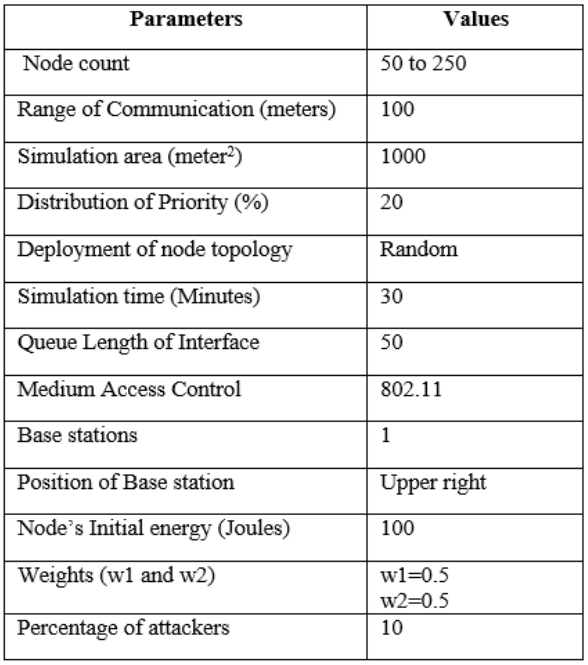
VI. Results

NS2 is used to simulate the proposed solution. The simulation was conducted with following parameters that are shown in Fig. 9.

The proposed solution is compared with solution proposed in [17] for ensuring survivability in contrast to black hole attacks and preserving energy efficiency and solution planned in [20] for localized secure routing architecture in contrast to cooperative black hole attacks. The performance is analyzed in terms of following criterion:

1) Node’s Life time.
2) Packet delivery ratio.
3) Node’s average energy.
4) End to End delay.
5) Histogram of energy.

Life time is the time at which the energy of the first node goes to 0. The life time is calculated for different number of nodes and is plotted in Fig. 10.



Parameters	Values
Node count	50 to 250
Range of Communication (meters)	100
Simulation area (meter^2)	1000
Distribution of Priority (%)	20
Deployment of node topology	Random
Simulation time (Minutes)	30
Queue Length of Interface	50
Medium Access Control	802.11
Base stations	1
Position of Base station	Upper right
Node’s Initial energy (Joules)	100

Weights (w1 and w2)	w1=0.5
	w2=0.5

Percentage of attackers	10

Fig. 9. Simulation Parameters.

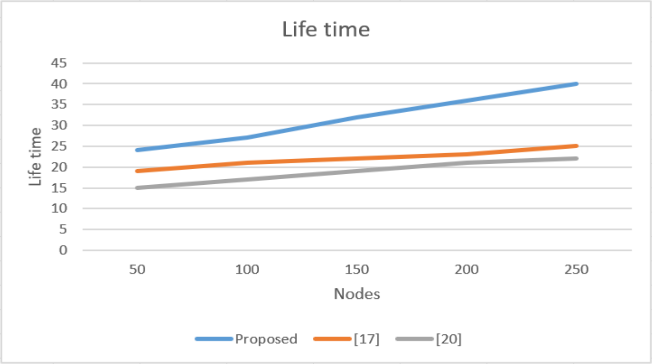


Fig. 10. Life Time Calculation.

The life time is more in the intended solution when compared to [17] and [20]. The life time is more in intended solution is due to because of efficient mitigation of attacks and energy balanced routing path selection. The life time comparison is shown in Table III.

The network packet delivery ratio is measured by changing the number of nodes in the network and the result is given Fig. 11.

The packet delivery ratio is more in the intended method compared to [17] and [20] due to resilient path selection in the proposed protocol. Table IV shows packet delivery ratio.

The packet delivery ratio is determined for different rate of packets and the result is given in Fig. 12.

The packet delivery ratio as shown in Table V, drops as the rate of packet increases, but the drop is very low in the proposed solution compared to [17] and [20]. The reason being the path selection procedure followed in the proposed protocol.

The energy at nodes is averaged at different interval of time and plotted in Fig. 13.

The average energy in the proposed solution (Table VI) is better than [17] and [20] due to well-balanced routing in the proposed solution.

TABLE III. LIFE TIME COMPARISON

No of Nodes	Proposed	[17]	[20]
50	24	19	15
100	27	21	17
150	32	22	19
200	36	23	21
250	40	25	22

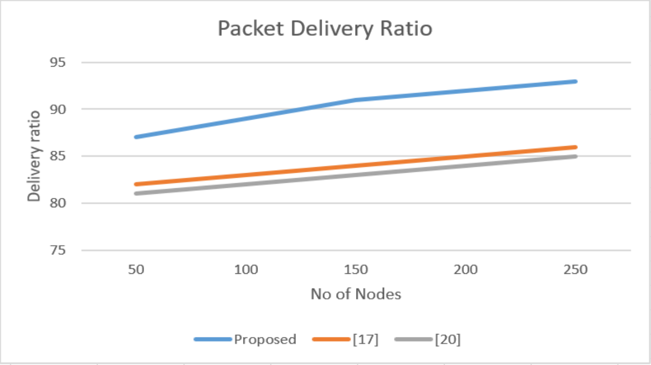


Fig. 11. Packet Delivery Ratio based on no.of Nodes.

TABLE IV. PACKET DELIVERY RATIO COMPARISON

No of Nodes	Proposed	[17]	[20]
50	87.4	82.15	81.5
100	89.3	83.34	82.17
150	91.1	84.56	83.11
200	92.5	85.12	84.32
250	93.3	86.31	85.5

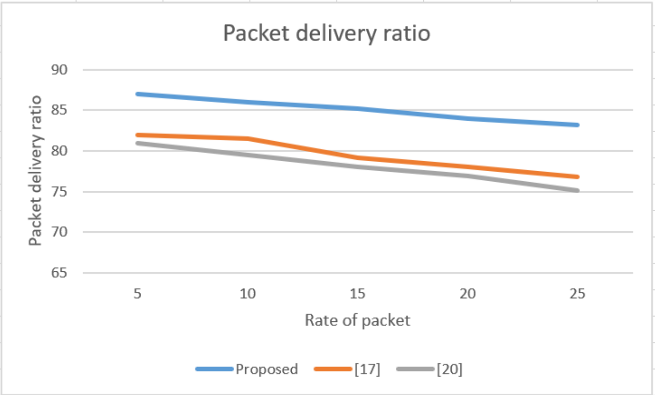


Fig. 12. Packet Delivery Ratio based on Rate of Packets.

TABLE V. PACKET DELIVERY RATIO

No of Nodes	Proposed	[17]	[20]
5	87	82	81
10	86	81.5	79.5
15	85.2	79.12	78
20	84	78	76.9
25	83.2	76.8	75.1

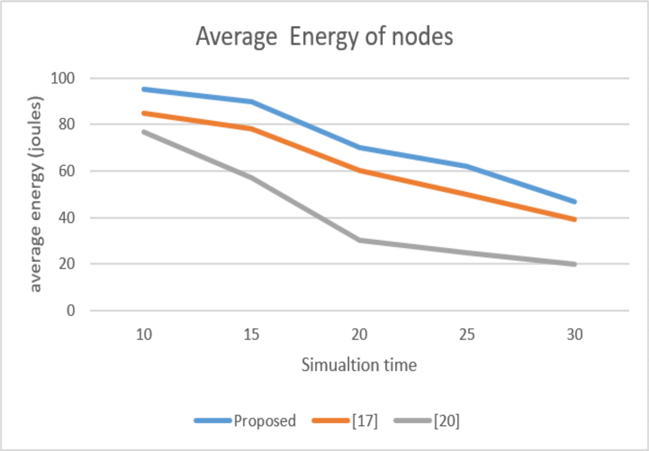


Fig. 13. Average Energy of Nodes.

TABLE VI. AVERAGE ENERGY OF NODES COMPARISON

Simulation time	Proposed	[17]	[20]
10	95	85	77
15	90	78	57
20	70	60	30
25	62	50	25
30	47	39	20

Average end to end delay for packet traversal from node to sink is derived and plotted in Fig. 14.

Due to offloading, all the computations to sink the delay for packet processing is reduced in nodes in the proposed solution compared to [17] and [20]. Delay comparison is represented in Table VII.

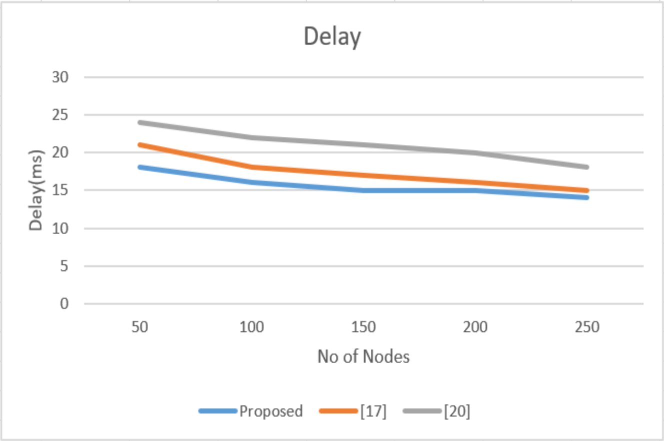


Fig. 14. Delay Comparison.

TABLE VII. DELAY COMPARISON

Simulation Time	Proposed	[17]	[20]
50	18	21	24
100	16	18	22
150	15	17	21
200	15	16	20
250	14	15	18

Histogram of energy is the distribution of nodes depending on their current residual energy. The total energy of 100 joules is split into 5 equal range and the number of nodes whose residual energy falling in the corresponding range is measured and histogram is plotted. The energy histogram gives an indication of how much duration the network will last or how many numbers of nodes are nearing their end of life time. Energy Histogram is shown in Fig. 15.

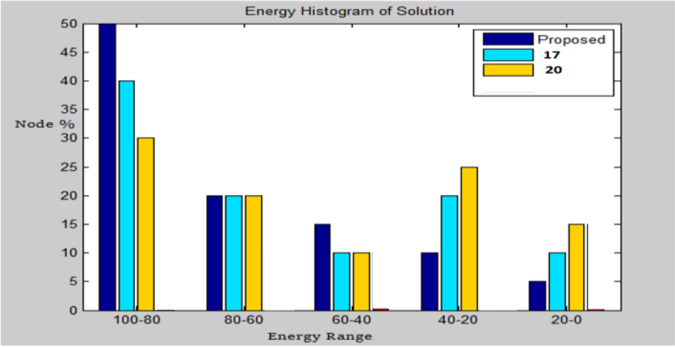


Fig. 15. Energy Histogram.

From the results, it can be seen that nodes with higher energy is available more in intended solution compared to [17] and [20] indicating prolonged life time in the intended solution.

VII. CONCLUSION

In this work, a secured energy efficient attack resilient routing protocol is planned for wireless sensor networks. The routing paths are adaptively scored using a fuzzy function on security and energy consumption. Instead of identifying malicious nodes individually, the proposed solution identified the security risks in a zone and proposed a mitigation mechanism to reduce the probability of routing in those less secure areas. The proposed solution is also able to select the routing path and prolong the life time of the network. The proposed solution can be extended to defend against some more attacks like worm hole as part of future work.

REFERENCES

[1] S. Rani, J. Malhotra, R. Talwar, "EEICCP-Energy Efficient Protocol for Wireless Sensor Networks", Wireless Sensor Network, vol. 5, no. 7, pp. 127-136, 2013.

[2] Nikolidakis, Stefanos A., et al. "Energy efficient routing in wireless sensor networks through balanced clustering." Algorithms 6.1 (2013): 29-42.

[3] Khan, Zahoor A., et al. "A QoS-aware routing protocol for reliability sensitive data in hospital body area networks." Procedia Computer Science 19 (2013): 171-179.

[4] Velasquez-Villada, Carlos, and Yezid Donoso. "Multipath routing network management protocol for resilient and energy efficient wireless sensor networks." Procedia Computer Science 17 (2013): 387-394.

[5] Ahvar, Ehsan, et al. "An energy-aware routing protocol for query-based applications in wireless sensor networks." The Scientific World Journal 2014 (2014).

[6] Monowar, Muhammad Mostafa. "An Energy-aware Multi-constrained Localized QoS Routing for Industrial Wireless Sensor Networks." Adhoc & Sensor Wireless Networks 36 (2017).

[7] Kim, Kyung Tae, and Hee Yong Youn. "An Energy-Efficient and Scalable Routing Protocol for Distributed Wireless Sensor Networks." Adhoc & Sensor Wireless Networks 29 (2015).

[8] Khodabandeh, Hajar, Vahid Ayatollahitafti, and Mohammad Sadeq Taghizadeh. "Link aware and Energy efficient Routing Algorithm in Wireless Body Area Networks." Netw. Protoc. Algorithms 9.1-2 (2017): 126-138.

[9] Mehmood, Amjad, et al. "ELDC: An artificial neural network based energy-efficient and robust routing scheme for pollution monitoring in WSNs." IEEE Transactions on Emerging Topics in Computing (2017).

[10] Manjeshwar, Arati, and Dharma P. Agrawal. "TEEN: A Routing Protocol for Enhanced Efficiency in Wireless Sensor Networks." ipdps. Vol. 1. 2001.

[11] Brar, Gurbinder Singh, et al. "Energy efficient direction-based PDORP routing protocol for WSN." IEEE access 4 (2016): 3182-3194.

[12] Gherbi, Chirihane, Zibouda Aliouat, and Mohammed Benmohammed. "Distributed energy efficient adaptive clustering protocol with data gathering for large scale wireless sensor networks." 2015 12th International Symposium on Programming and Systems (ISPS). IEEE, 2015.

[13] Kandukuri, Somasekhar, Nour Murad, and Richard Lorion. "A single-hop clustering and energy efficient protocol for wireless sensor networks." 2015 Radio and Antenna Days of the Indian Ocean (RADIO). IEEE, 2015.

[14] Kandukuri, Somasekhar, et al. "Energy-efficient data aggregation techniques for exploiting spatio-temporal correlations in wireless sensor networks." 2016 Wireless Telecommunications Symposium (WTS). IEEE, 2016.

[15] Robinson, Y. Harold, and M. Rajaram. "Energy-aware multipath routing scheme based on particle swarm optimization in mobile ad hoc networks." The Scientific World Journal 2015 (2015).

[16] Jhaveri, Rutvij H., and Narendra M. Patel. "A sequence number based bait detection scheme to thwart grayhole attack in mobile ad hoc networks." Wireless Networks 21.8 (2015): 2781-2798.

[17] Khamayseh, Yaser M., Shadi A. Aljawarneh, and Alaa Ebrahim Asaad. "Ensuring survivability against Black Hole Attacks in MANETS for preserving energy efficiency." Sustainable Computing: Informatics and Systems 18 (2018): 90-100.

[18] Khatoon, Naghma. "Mobility aware energy efficient clustering for MANET: a bio-inspired approach with particle swarm optimization."

Wireless Communications and Mobile Computing 2017 (2017).

[19] Nallusamy, C., and A. Sabari. "Particle Swarm Based Resource Optimized Geographic Routing for Improved Network Lifetime in MANET." Mobile Networks and Applications 24.2 (2019): 375-385.

[20] Poongodi, Thangamuthu, and M. Karthikeyan. "Localized secure routing architecture against cooperative black hole attack in mobile ad hoc networks." Wireless Personal Communications 90.2 (2016): 1039-1050.

[21] Rajan, C., et al. "Investigation on novel based naturally-inspired swarm intelligence algorithms for optimization problems in mobile ad hoc networks." World Academy of Science, Engineering and Technology International Journal of Mathematical, Computational, Natural and Physical Engineering 9.3 (2015).

[22] Kadiravan, G., Pothula Sujatha, and J. Amudhavel. "A state of art approaches on energy efficient routing protocols in mobile wireless sensor networks." IIOAB JOURNAL 8.2 (2017): 234-238. www. https://www.iioab.org/

[23] Yan, Ziwei, et al. "Energy-efficient node positioning in optical wireless sensor networks." Optik 178 (2019): 461-466.

[24] Arage Chetan, S., & Satyanarayana, K. V. V. “Novel routing protocol for secure data transmission in wireless ad hoc networks.” International Journal of Innovative Technology and Exploring Engineering, 8(4S2) (2019)., 101-108.

[25] Karthikeyan, T., V. Brindha, and P. Manimegalai. "Investigation on Maximizing Packet Delivery Rate in WSN Using Cluster Approach." Wireless Personal Communications 103.4 (2018): 3025-3039.

[26] Chowdary, Krishna, and K. V. V. Satyanarayana. "Malicious Node Detection and Reconstruction of Network in Sensor Actor Network." Journal of Theoretical & Applied Information Technology 95.3 (2017).

[27] Gummadi, Annapurna, and K. Raghava Rao. "EECLA: Clustering And Localization Techniques To Improve Energy Efficient Routing In Wireless Sensor Networks." Journal of Theoretical & Applied Information Technology 96.1 (2018).

[28] Vamshi krishna, H., & Swain, G. “Identification and avoidance of malicious nodes by using certificate revocation method.” International Journal of Engineering and Technology(UAE), 7(4.7 Special Issue 7) (2018)., 152-156.

[29] Mallikarjuna Rao, Y., M. V. Subramanyam, and K. Satya Prasad. "Cluster-based mobility management algorithms for wireless mesh networks." International Journal of Communication Systems 31.11 (2018): e3595.

[30] Chowdary ,K., & Satyanarayana, K. V. V. “A novel secured data transmission and authentication technique against malicious attacks in WSNs.” Journal of Advanced Research in Dynamical and Control Systems, (Special Issue -18) (2017), 161-173.

[31] Kalaipriyan, T., et al. "Monkey King Algorithm for Solving Minimum Energy Broadcast in Wireless Sensor Network." Advances and Applications in Mathematical Sciences 17.1 (2017): 129-145.

[32] Dr. P.V. Rao, Manjunath B E, “Unique Analytical Modelling of Secure Communication in Wireless Sensor Network to Resist Maximum Threats”, International Journal of Advanced Computer Science and Applications, (IJACSA) Vol. 10, No. 2, pp 421-427, 2019.

[33] Blanza, F., and L. Materum. "Joint Identification of the Clustering and Cardinality of Wireless Propagation Multipaths." International Journal of Emerging Trends in Engineering Research 7.12 (2019): 762-767.

[34] Al-Shargabi, Bassam, and Mohammed Aleswid. "Performance of RPL in Healthcare Wireless Sensor Network." International Journal of Emerging Trends in Engineering Research 8.3 (2020).

An Effective Solution to Count-to-Infinity Problem for both Complex and Linear Sub-Networks

Sabrina Hossain[1], Kazi Mushfiquer Rahman[2], Ahmed Omar[3], Anisur Rahman[4]
Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh

***Abstract*—Distance vector routing protocol determines the best route for forwarding information from one node to another node based on distance. For calculating the best route, Distance-vector routing protocols use the Bellman-ford algorithm and the Ford-Fulkerson algorithm. The Bellman-Ford distributed algorithm calculates the shortest path. On the other hand, Routing Information Protocol is commonly used for managing router information management protocol within a Local Area Network or an interconnected Local Area Network group. The main problem with Distance Vector Routing protocols is routing loops. Because the Bellman-Ford Algorithm cannot prevent loops. Moreover, the routing loop triggers a problem with Count to Infinity. This research paper gives an effective solution to the Count to Infinity problem for link down situation and also for the router down situation in both complex and linear sub-network. For the router down situation, when any router goes down, then other nodes will recalculate their routing table with the dependency column. Moreover, the costs are calculated by the shortest path algorithm. If any link is down and all routers are up, then all routers will recalculate their routing table using Dijkstra instead of the Bellman-Ford algorithm. To determine the loops and prevent the loops are the main objectives. This method is mainly based on a routing table algorithm where the Dijkstra algorithm will be used after each iteration and will modify the routing table for each node. Preventing the routing loops will not converge into Count to Infinity Problem.**

Keywords—Distance vector routing; local area network; routing information protocol

I. INTRODUCTION

Nowadays, scientists are trying to reduce the packet loss problem. Everybody needs high-speed internet and, they want everything fast. However, because of packet loss, the service becomes slow, the network connection gets disrupted, and sometimes it loses the whole network connectivity. It creates significant problems in real-time data transfer programs. So, a better network means less packet loss. Distance Vector Routing is one of the dynamic algorithms [1]. Whenever a router goes down, routing loops usually occur in DVR. Then it creates a linear topology. Linear topology generates count to infinity problem. Because of that, a huge number of packets will be lost and, that is an immense issue. If all router gets to know earlier that any router already got down by observing an extra column of the routing table, then the packet loss problem can be minimized and which will improve the network connectivity. The modification of the routing table and alternation of the algorithm into the Dijkstra algorithm might cause less packet loss problem, and then the network connectivity will become well.

II. ROUTING ALGORITHM

Routing is a method of determining the routes to reach the destination that data packets will obey. A table of routing table is created in this process which contains information about the routes that data packets follow. Different routing algorithms are used to determine which route an incoming data packet needs to be transmitted efficiently to its destination.

A. Distance Vector Routing Algorithm

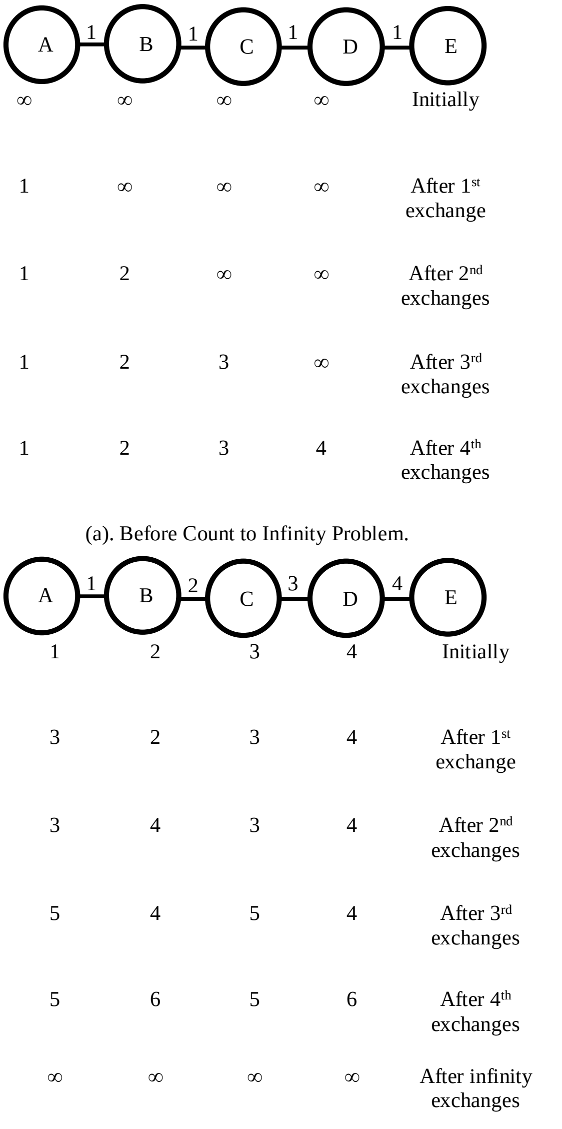
The Distance vector uses the Bellman-Ford algorithm for finding the shortest paths [2]. It can also be calculated by Dijkstra algorithm. Every node calculates the distance from other routers. The shortest path is created based on the metric. The metric is referred to as a count or a distance. In the Distance vector, the process of exchanging information is done iteratively [1]. There is no information exchange between the neighbourhoods until the information received from at least one neighbour directly and the algorithm does not require all the neighbours are asynchronous with each other. In distance vector, each node maintains the distance from it, to its possible destination and sends a periodic routing update. For periodic routing updates, the convergence time is slow. The slow convergence leads to count-to-infinity and routing loops problem [1].

B. Dijkstra Algorithm

Dijkstra algorithm solves the shortest path algorithm, and it is better than the Bellman-Ford algorithm. It works better when multiple paths present in the topology, and it helps to choose the shortest path [3][4]. In the following algorithm, the code u ← vertex in Q with min dist[u], searches for the vertex u in the vertex set Q that has the least dist[u] value. Length (u, v) returns the length of the edge joining (i.e. the distance between) the two neighbour-nodes u and v. The variable alt on line 18 is the length of the path from the root node to the neighbour node v if it were to go through u. If this path is shorter than the current shortest path recorded for v, that current path is replaced with this alt path. The previous array is populated with a pointer to the "next-hop" node on the source graph to get the shortest route to the source [5]. The Dijkstra algorithm is a Dynamic programming approach. A complex problem is divided into sub-problems in the Dynamic programming approach, then combine the solutions of these sub-problems to get an overall solution. So, Dijkstra's algorithm is a greedy approach. So, it does not create routing loops. In the Bellman-Ford algorithm, it cannot prevent loops, but through Dijkstra, it can be prevented. If any link or router down then, it will apply a greedy approach that will prevent the loops and solve the Count to infinity problem.

III. Count to Infinity Problem

Count-to-Infinity Problem is one of the most important issues in Distance Vector Routing (DVR) Algorithm. When an interface goes down, routing loops usually occur in DVR. Which actually creates linear subnet. When two routers send an update to each other at the same time, it can also occur [6][7]. Distance Vector Routing reacts rapidly to good news, but leisurely to bad news [8]. In distance vector routing, it uses the Bellman-Ford algorithm to propagate. To see how slow bad news propagates, consider the situation of 1(a) in which all the lines and router are initially up. Router B, C, D, and E have the distance to A of 1, 2, 3, and 4 respectively. Suddenly A goes down, or the line between A and B is cut, which is effectively the same thing from B's point of view in Fig. 1(b). At the first packet exchange, B does not hear anything from A. Fortunately, C says: do not worry; I have a path to A of length 2. As a result, B thinks it can reach A via C, with a path length of 3 and D & E do not update their entries for A on their first exchange.



∞	∞	∞	∞	Initially
1	∞	∞	∞	After 1st exchange
1	2	∞	∞	After 2nd exchanges
1	2	3	∞	After 3rd exchanges
1	2	3	4	After 4th exchanges

(a). Before Count to Infinity Problem.

1	2	3	4	Initially
3	2	3	4	After 1st exchange
3	4	3	4	After 2nd exchanges
5	4	5	4	After 3rd exchanges
5	6	5	6	After 4th exchanges
∞	∞	∞	∞	After infinity exchanges

Fig. 1. (b). After Count to Infinity Problem.

On the second exchange, C notices that each of its neighbour's claims to have a path to A of length 3. It picks one of them at random and makes its new distance to A of length 4, as shown in the third row. Subsequent exchanges produce the history shown in the rest of Fig. 1(a) and 1(b) [8].

Linear Subnet: Whenever a router goes down, routing loops usually occur in DVR. Then it creates a linear topology. Linear topology creates count to infinity problem.

In Fig. 2 a complex network is shown, but every time a router goes down, the routing loops will happen.

If router E from Fig. 2 goes down, then there will be a linear sub-network which is shown in Fig. 3.

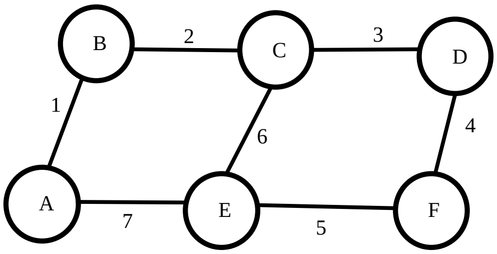


Fig. 2. A Complex Network.

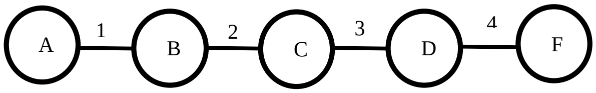


Fig. 3. A Linear Sub-Network.

IV. Related Work

The main difference between link sate and distance vector routing is, link-state uses an algorithm derived from Dijkstra's shortest path algorithm where distance-vector uses a distributed Bellman-Ford algorithm [8]. The distance vector routing algorithm suffers from the count-to-infinity problem. Count-to-infinity problems can be solved by preventing loops [9][8][1]. Researchers tried to prevent them by using various ways. Mr. D. Ganesh solved the count to infinity problem by using RSTP (Rapid Spanning Tree) protocol, Bridge protocol unit, building and maintaining SP (Spanning Tree) tree and changing the topology. They also follow some rules which are If a bridge can no longer reach the root bridge via its root port and does not have an alternate port, it declares itself to the root; A bridge sends out a BPDU immediately after the topology information it is announcing has changed, A designated port becomes the root port if it receives a better BPDO than what the bridge has received before. That is, this BPDU announces a better path to the root than via the current root port. When a bridge loses connectivity to the root bridge via its root port, and it has one or more alternative ports, it as its new root port. By using these rules, they portioned the network. Whenever a network is partitioned, if the partition does not contain the root bridge as a cycle, there exists a race condition that can result in the count-to-infinity behaviour. Count-to-infinity may even occur without a network partition. This new topology information will go around the loop until it reaches an alternate port caching stale, but better information. Again this stale information will choose the new information around the loop in a count to infinity. This will keep going until the stale topology information reaches its massage [8]. Amit D. Kothari and Dharmendra T. Patel have also solved the count-to-infinity problem by using the test packet. They have some criteria for this test packet. First of all, they give source and destination addresses with the sequence number. They also declare the packet type like a query for any router and answer for the router. They also count the hop which initializes with 0 and increments by each intermediate router. The test packet will travel through source address to lastly. Then the address of the test router via the test packet is to be forwarded. Then it will give status where 1 is positive, and 0 is negative. The value will be a delay for the last neighbour. For error handling it will checksum. To solve the count to infinity, they design this type of test packet [1].

There is various way to solve the Count to Infinity.

1) Routing information protocol uses split horizon. Split horizon is a process where the actual distance to a destination is not reported to a node through which reaches the destination For example if node A has learned a route to node C through B, then A does not send the distance vector of C to node B during a routing update [9].

2) The count to infinity problem can be avoided by using hold-down timers. This is a clock that is set within the node to help ensure network stability.

When a node receives an update from a neighbour indicating that a previously accessible network is not working and is inaccessible, the hold-down timer will start. If a new update arrives from a neighbour with a better metric than the original network entry, the hold-down is removed, and data is passed. Nevertheless, an update is received from a neighbour node before the hold-down timer expires and it has a lower metric than the previous route. Therefore the update is ignored, and the hold-down timer keeps ticking. This allows more time for the network to converge [9].

V. Methodology

A. Procedure

Aiming at the difficulties in the count to infinity problem, this research proposed a method using Dijkstra Algorithm in each iteration and modifying the routing table with additional information. In the routing table, it added cost (shortest path), dependency, and status.

1) Dependency: It denotes the dependency of a router. For example, for calculating the routing table for C if it chooses a path from Router A to C via B and when the node A is down from Fig. 4 then, B is dependency router of C.

2) Status: It denotes the router is in hold situation or not. Which will prevent the routed loop in DVR.

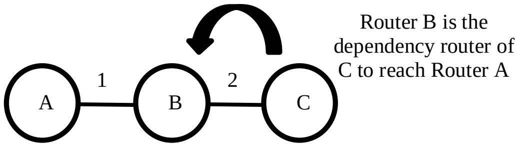


Fig. 4. A Linear Sub-Network of Three Routers.

This research has designed a method for both complex network and linear sub-network. At first, it will check if there is any link down or not. However, it has four different cases.

Case1: If all links are up, then it will go for checking the routers. If any routers are not down, then every node will calculate their routing table.

Case2: If all links are up, but any of the routers is down, then every router will recalculate their routing table. Every router will check the dependency column while recalculating its routing table. From the dependency column, a router can get to know about the present status of each router. Which means a router can easily track if there is any router down or not. If a router finds any router as down and which is its dependency router, then it will also change its status as down. Gradually all router's status will be down, and it will not create count to infinity problem.

Case3: If any link is down and all routers are up, then all routers will also recalculate their routing table using Dijkstra. While recalculating the routing table, a router will call itself as a holding router by updating the status column in the routing table. Therefore, other routers can do their work without any obstacles, and in the meantime, they can transmit their required packets to each other.

Case4: If any link is down and any router is also down then, at first it will follow case number 3 as mentioned in the above. And then it will follow the case of router down.

B. Algorithm of Proposed Method

In the two situations, one is for the topology links, and another is for routers. In four cases, two of the cases occur for the link down situation, and another two cases occur for router down situation.

```
1: Function main(){
Randomly pick up any cases
2: If (case 1){
Run generate routing table (source)
Print routing table
}
3: If (case 2){
 I. Scan which router is down
 II. Switch down the router
 III. Run generate routing table (source)
 IV. Print routing table
}
4: If (case 3){
 I. Scan which link is down
          II. Turn off the link
          III. Run generate routing table (any node of the link)
          IV. Print distance between the link's node
}
5: If (case 4){
 I. Scan the link which is down
          II. Turn off the link
          III. Calculate distance vector
          IV. Print distance vector
}
}
```

As the shortest path algorithm, it will use the Dijkstra Algorithm. Which is,

```
Function generate routing table(source){
    create vertex set Q
    for each vertex v in Graph:
        dist [v] ← INFINITY
        prev[v] ← UNDEFINED
        add v to Q
    dist[source] ← 0
    while Q is not empty:
        u ← vertex in Q with min dist[u]
        remove u from Q
        for each neighbor v of u: // only v that are still in Q
            alt ← dist[u] + length(u, v)
            if alt < dist[v]:
                dist[v] ← alt
                prev[v] ← u
    return dist[], prev[]
}
```

For preventing the routing loops if all routers state their dependency with the additional details in the routing table, then all nodes will be slowly informed if there is any router in the down condition which causes routing loops. Here, the routing loops problem will be prevented. Moreover, for link down situation packet loss problem may be arrived. So, if any link is down, then all routers will recalculate their routing table using Dijkstra. While recalculating the routing table, a router will call itself as a holding router by updating the status column in the routing table.

VI. Simulation Result

A. Simulation Result for Complex Network

The proposed method has been simulated using C++ to validate the proposed model. A group of four routers are placed at different costs, see Table I. Fig. 5 of the graph is given below.

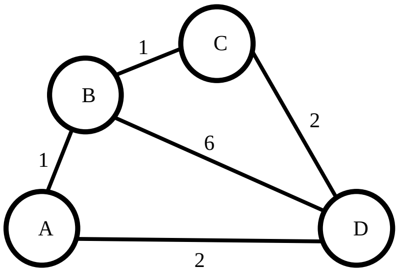


Fig. 5. A Graph with Various Costs.

TABLE I. The Costs of the paths

Routers	A	B	C	D
A		1		2
B	1		1	6
C		1		2
D	2	6	2	

At first, it will check if there is any link down or not. Then it will choose a case randomly from four cases. If it chooses Case 1 when the source router is A, then the routing Table II is going to appear as mentioned.

TABLE II. Case 1 with source router A

Node	Cost	Dependency	Status
A	0	NULL	UP
B	1	A	Up
C	2	B	Up
D	4	C	Up

If it chooses Case 2 when the source router is A, and it goes down, then the routing Table III will be like this

TABLE III. Case 2 with source router A

Node	Cost	Dependency	Status
A	0	NULL	Down
B	1	A	Up
C	2	B	Up
D	4	C	Up

Then the router B's status will be down as its dependency router A's status is also down in Table IV.

TABLE IV. Router B's status is down when its dependency router's status A is also down

Node	Cost	Dependency	Status
A	0	NULL	Down
B	1	A	Down
C	2	B	Up
D	4	C	Up

Then the router C's status will be down as its dependency router B's status is also down in Table V.

TABLE V. Router C's status is down when its dependency router B's status is also down

Node	Cost	Dependency	Status
A	0	NULL	Down
B	1	A	Down
C	2	B	Down
D	4	C	Up

At last the router D's status will be down as its dependency router C's status is also down in Table VI.

TABLE VI. Router D's status is down when its dependency router C's status is also down

Node	Cost	Dependency	Status
A	0	NULL	Down
B	1	A	Down
C	2	B	Down
D	4	C	Down

So it will not cause the count to infinity problem. This is how this method handles Count to Infinity problem for a complex network.

If it chooses Case 3 when the source is A, and the link between A and B is down, then the graph will be appear as Fig. 6.

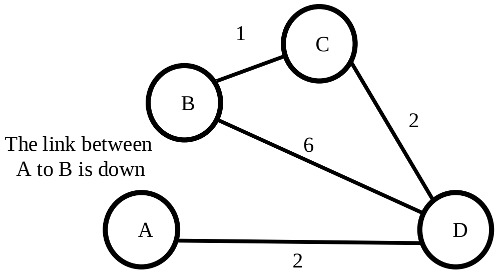


Fig. 6. The Link between A and B is Down.

Then it will recalculate the shortest path with the Dijkstra Algorithm. And then the new path from router A to router B will be **A->D->C->B** which is shown in the Fig. 7.

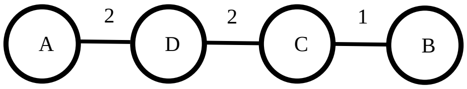


Fig. 7. New Path from Router A to Router B.

If it chooses case 4, then at first, it will follow case number 3 as mentioned in the above. And then it will follow the case of router down.

B. Simulation Result for Linear Sub-Network

Count to Infinity problem occurs due to linear sub-network. In Count to Infinity Problem, the updating of the routing table continues infinite time.

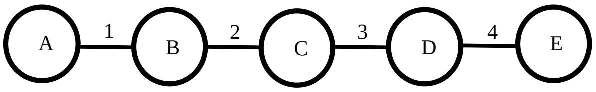


Fig. 8. A Linear Sub-Network of Five Routers.

In the linear sub-network of Fig. 8, when router A goes down, then the first five exchange of information is going to appear as mentioned in Table VII.

After 101 number of exchange of information, the routing Table VIII will be like this.

TABLE VII. FIRST FIVE EXCHANGE OF INFORMATION AFTER ROUTER A IS DOWN

Router	**A**	**B**	**C**	**D**	**E**
1st Exchange	NULL	5	3	6	10
2nd Exchange	NULL	5	7	6	10
3rd Exchange	NULL	9	7	10	10
4th Exchange	NULL	9	11	10	14
5th Exchange	NULL	13	11	14	14

TABLE VIII. AFTER 101ST NUMBER OF EXCHANGE OF INFORMATION

Router	**A**	**B**	**C**	**D**	**E**
102nd Exchange	NULL	205	207	206	210
103rd Exchange	NULL	209	207	210	210
104th Exchange	NULL	209	211	210	214
105th Exchange	NULL	213	211	214	214

Moreover, the routing Table IX will be updated for an infinite time. So, this is how the count to infinity problem occurs in linear sub-network. Nevertheless, in this proposed method, when router A goes down, then it will gradually let other routers know about its (Router A) condition. So, Count to infinity will not occur. If this proposed method applied for Fig. 8 when router A goes down, then this method will give a solution like this.

TABLE IX. AFTER ROUTER A GOES DOWN

Node	**Cost**	**Dependency**	**Status**
A	0	NULL	Down
B	1	A	Up
C	3	B	Up
D	6	C	Up
E	10	D	Up

Then the router B's status will be down as its dependency router A's status is also down in Table X.

TABLE X. ROUTER B'S STATUS IS DOWN AND ALSO IT'S DEPENDENCY ROUTER A'S STATUS IS DOWN

Node	**Cost**	**Dependency**	**Status**
A	0	NULL	Down
B	1	A	Down
C	3	B	Up
D	6	C	Up
E	10	D	Up

Then the router C's status will be down as its dependency router B's status is also down in Table XI.

Then the router D's status will be down as its dependency router C's status is also down in Table XII.

TABLE XI. ROUTER C'S STATUS IS DOWN AND ALSO IT'S DEPENDENCY ROUTER B'S STATUS IS DOWN

Node	**Cost**	**Dependency**	**Status**
A	0	NULL	Down
B	1	A	Down
C	3	B	Down
D	6	C	Up
E	10	D	Up

TABLE XII. ROUTER D'S STATUS IS DOWN AND ALSO IT'S DEPENDENCY ROUTER C'S STATUS IS DOWN

Node	Cost	Dependency	Status
A	0	NULL	Down
B	1	A	Down
C	3	B	Down
D	6	C	Down
E	10	D	Up

At last the router E's status will be down as its dependency router D's status is also down in Table XIII.

TABLE XIII. ROUTER E'S STATUS IS DOWN AND ALSO IT'S DEPENDENCY ROUTER D'S STATUS IS DOWN

Node	Cost	Dependency	Status
A	0	NULL	Down
B	1	A	Down
C	3	B	Down
D	6	C	Down
E	10	D	Down

So, in this method, there will be no routing loops. So, the Count to Infinity Problem will not occur.

C. Simulation of Graphs

Generally, Count to Infinity occurs in linear sub-network. All routers of the topology gradually increase their routing table for an infinite time. A graph of Count to Infinity is given.

From Fig. 9 there is a combined graph of 4 routers of Fig. 8. In Fig. 9, where the blue curve is for Router B. Before router A goes down the cost of B was 1. However, after router A got down, then router B updated its cost with the help of router C. Nevertheless, C's cost was updated with the help of router B before. So, after the 1st exchange of information, the cost of router B was updated from 1 to 5. Moreover, it will gradually update its cost for an infinite time. Furthermore, all routers followed the way of updating the cost as router B followed. So, all curves of the routers converge to infinity. In Fig. 9, the 1st – 5th exchanges, 102nd – 105th exchanges, and 1002nd – 1005th exchanges have shown.

If each Router's costs plot respect to the number of exchanges in this method, then a graph as Fig. 10 will appear.

According to Fig. 8, the cost of router B was 1 before router A goes down. However, in the 1st exchange of information, the cost of the router will not change as there was direct dependency with router A, but now it is down. After 2nd exchange of information, router B's status will also be down as its dependency router A's status is already down. In 3rd exchange of information when the router C wants to update its cost its status will be down as its dependency router B's status is already down. Gradually router D and E will change their status as down while checking the dependency. After 4th exchange of information, all router's status will be down. So the curves of all routers will not converge to infinity.

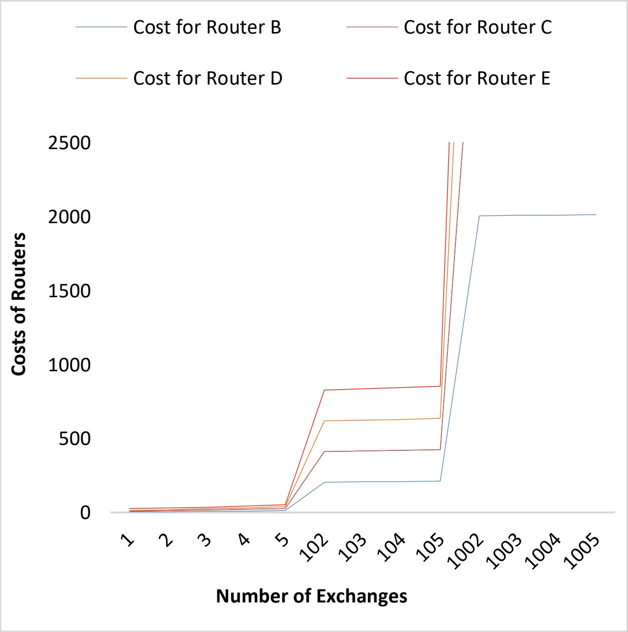


Fig. 9. Number of Exchanges vs each Router's Costs for Count to Infinity.

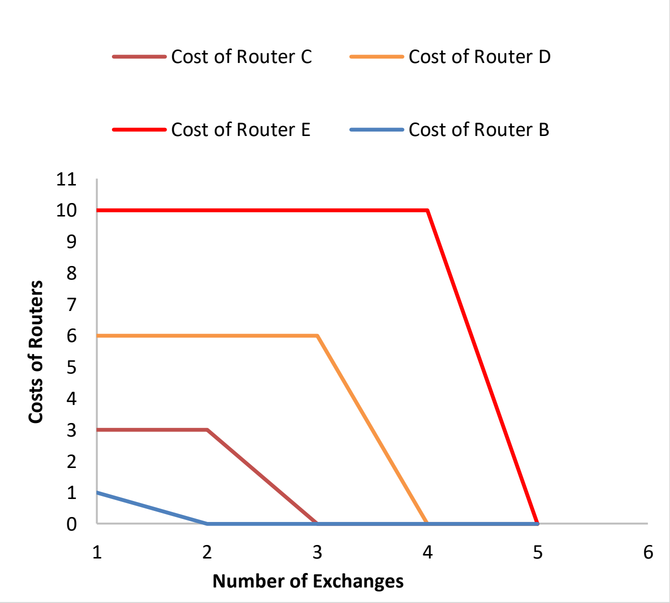


Fig. 10. Number of Exchanges vs each Router's Costs for Proposed Method.

D. Comparison of Complexity

Thus the count to infinity problem uses the Bellman-Ford algorithm, so the time complexity of Count to infinity is O(V^2). However, this method implemented Dijkstra instead of the Bellman-Ford algorithm. Though, Bellman-Ford is simpler than Dijkstra and suites well for distributed systems. Nevertheless, the time complexity of Bellman-Ford is more than Dijkstra. In this method, the Dijkstra Algorithm with minimum priority queue can be reduced the complexity to O(V + E log V).

VII. Conclusion

For avoiding the Count-To-Infinity problem and reducing the packet loss, this method is applying the Dijkstra algorithm instead of the Bellman-Ford algorithm and that solve the Count-To-Infinity problem and reduce the packet loss. Now it does not face any problem with real-time data transfer, and the network connection will be undisrupted. This problem is solved for two types of situations one is router down, and the other one is linked down. For router down, the method handles the situation by giving additional information about the router, which is dependency. When any router goes down, then other nodes will recalculate their routing table with the dependency column. The shortest path algorithm calculates the costs. For link down situation, if any link is down and all routers are up, then all routers will recalculate their routing table using the Dijkstra algorithm instead of the Bellman-Ford algorithm. These are an effective way to solve the Count-To-infinity problem. There are some mechanisms known, such as defining the maximum count, split horizon, poison reverse, triggered update, and hold down timer. However, using this proposed way, an effective result will come out, and the packet loss will become less.

References

[1] D. Kothari and D. T. Patel, "Methodology to Solve the Count-To-Infinity Problem by Accepting and Forwarding Correct and Updated Information Only Using Test Packet," 2009 IEEE Int. Adv. Comput. Conf. IACC 2009, no. April, pp. 26–31, 2009, doi: 10.1109/IADCC.2009.4808974.

[2] Alberto Leon-Garcia and India Widjaja. Communication Networks, Fundamental Concepts and Key Architectures. McGraw ́-]Hill Higher Education, Singapore, International Editions 2000. ISBN 0-07-022839-6.

[3] B. F. Zhan, "Three fastest shortest path algorithms on real road networks: Data structures and procedures," J. Geogr. Inf. Decis. Anal., vol. 1, no. 1, pp. 70–82, 1997.

[4] A. Goldberg and R. E. Tarjan, "Expected Performance of Dijkstra's Shortest Path Algorithm," Networks, no. 2 43, pp. 4–10, 1996.

[5] M. J. Bannister and D. Eppstein, "Randomized speedup of the Bellman-Ford algorithm," 9th Meet. Anal. Algorithmics Comb. 2012, ANALCO 2012, pp. 41–47, 2012, doi: 10.1137/1.9781611973020.6.

[6] K. Elmeleegy, A. L. Cox, and T. S. E. Ng, "Understanding and mitigating the effects of count to infinity in ethernet networks," IEEE/ACM Trans. Netw., vol. 17, no. 1, pp. 186–199, 2009, doi: 10.1109/TNET.2008.920874.

[7] K. Elmeleegy, A. L. Cox, and T. S. E. Ng, "On count-to-infinity induced forwarding loops in ethernet networks," Proc. - IEEE INFOCOM, 2006, doi: 10.1109/INFOCOM.2006.229.

[8] V. Rama and P. Vaddella, "An Effective Solution to Reduce Count-to-Infinity Problem in Ethernet," Int. J. Comput. Sci. Issues, vol. 7, no. 4, pp. 44–49, 2010.

[9] R. K. MCA and R. U. MCA, "an Exploration of Count-To-Infinity Problem in Networks," Ijest.Info, vol. 2, no. 12, pp. 7155–7159, 2010, [Online]. Available: http://www.ijest.info/docs/IJEST10-02-12-065.pdf.

An Efficient Smart Weighted and Neighborhood-enabled Load Balancing Scheme for Constraint Oriented Networks

Mohammed Amin Almaiah
Department of Computer Networks and Communications
College of Computer Sciences and Information Technology, King Faisal University, Saudi Arabia

***Abstract*—In Wireless Sensor Networks (WSNs), uniform load or traffic distribution strategy is one of the main challenging issues, which is tightly coupled with the resource-limited networks. To address this problem, various mechanisms have been developed and presented in the literature. However, these approaches were either application specific that is designed for a specific application area such as smart building or overlay complex. Therefore, a simplified and energy efficient load-balancing scheme is always needed for the resource-limited networks. In this paper, an efficient and neighborhood-enabled load-balancing scheme is presented to resolve the aforementioned issues specifically with available resources. For this purpose, the proposed scheme bounds every member node to collect various information about neighboring nodes i.e., those nodes resides in its communication range. Moreover, if residual energy *Er* of sensor node is less than the defined threshold value then it shares this information with neighboring nodes. In the proposed neighborhood-enabled load balancing scheme, every sensor node Ci prefers to route packets through the optimal paths particularly those paths where probability of critical nodes is negligible i.e., path where critical node(s) are not deployed. Simulation results showed that the proposed neighborhood-enabled load-balancing scheme is better than existing approaches in terms of network lifetime (both individual node and whole WSNs), throughput, and average packet delivery ratio and end-to-end delay performance metrics.**

Keywords—Wireless Sensor Networks (WSNs); load balancing; PSO; routing protocol; low power devices

I. Introduction

With the development of wireless technologies, Wireless Sensor Networks (WSNs) must have low cost, low power consumption, small size, easy layout, and dynamically self-organized [1]. WSNs are using widely in many different applications, such as industrial applications, health care, military defense, environmental detection, target tracking, and ecological observation, etc. [2, 3]. However, because batteries supply the energy of the sensors, its radio transmission distance will be limited. In order to solve the energy consumption and distance problems during transmission, if the sensors are too far away from the base station, the WSNs need to organize a large number of distributed nodes to form a multi-hop wireless network. In WSNS, the sensor uses multiple-hop to establish network routing. The data is transmitted back to the base station through the shortest path composed of multiple sensors [4]. In addition, in a WSN since the wireless sensor is powered by a battery, an energy efficient and reliable load balancing is relatively important. There are many approaches to reduce energy consumption in WSNs [5, 10]. Such as load balancing among wireless sensor nodes, path selection when transmitting data, scheduling problems, sleep mechanism when there is no sensitive data, correctness of transmitting data, and so on [11,34,35].

Generally, devices *Ci* belongs to the Wireless Sensor Networks (WSNs) have the capacity to organize them- selves to form an operational network without human intervention specifically in remote areas. These networks consist sensor nodes *Ci*, base station or sink node(s) *Sj*, actuator (where needed) and servers etc. These devices *Ci* are either deployed manually (known as engineered deployment) or randomly (usually through a helicopter or other means specifically in hard areas where manual deployment is not possible) in vicinity of the underlying phenomenon needed to be observed periodically or when a particular event occurs [1]. Due to the overwhelming characteristics of these networks, WSNs are used in different application areas such as military, smart-buildings, Medicine, agriculture, etc. to assist human beings in various real life-activities or automate it completely where applicable. Every member device *Ci*, i.e., sensor node, prob the environment periodically or when an event is triggered and shares the collected data with the cluster head (CH) or base station(s) or server(s) through a reliable communication mechanism that is either single-hop (in case of hierarchical WSNs) or multi hop (flat WSNs) [2]. In case of hierarchical networking infrastructures, wireless communication among various devices *Ci* is very simple as, usually, every device *Ci* is deployed within wireless communication range of the concerned base station module *Sj* [3]. Alternatively, in later case, each device *Ci* has to use path or rout information, i.e., load balancing or routing table, for the reliable transmission of data and it is due to the fact that majority of these devices are deployed far away from the concern base station and are unable to communicate directly [4,33]. Multi-hop communication is an ideal mechanism to resolve challenge of limited wireless communication that is closed related to the resource limited member nodes of operational WSNs. In this regard, various load balancing and routing schemes have been presented in literature to enhance the operational capabilities of these devices *Ci* specifically in terms of communication range, networks lifetime and end-to- end delay.

Shortest path with minimum cost path enabled routing schemes are considered among the most reliable communication techniques. Specifically for those systems that consider timely delivery of data has more priority than connectivity of networks like intruder or malicious detection. In these schemes, shortest path are identified and every member node C_i $WSNs$ is force to transmit data only through those path(s). One of the common problems associated with shortest path enabled schemes shortest lifespan of device C_i reside on the these shortest path which become the worst if multiple source nodes or devices share this shortest path for the transmission of packets [5], [6]. Hence, for the networks with limited resources, a uniform load balancing or traffic distribution strategy is always a challenging task specifically for the research community to resolve the aforementioned issue particularly with available resources. A well-known methodology that is multiple-path enabled load balancing schemes, which are primarily based on residual energy E_r of neighboring devices C_i, were presented to distribute net- work traffic uniformly across different communication paths [7]–[10]. In this scheme, sensor nodes can send data to the neighbors whose hop-count values are greater than the sender node. However, end-to-end delay and packets delivery ratio are among the common problems associated with these schemes as maximum packets are transmitted via longest paths available in the operational WSNs. Likewise, criticality aware load balancing schemes was proposed in literature, which has resolved some of the afore mentioned issues associated with multiple path enabled schemes such as net- works lifetime, average packet delivery ratio and throughput [11], [12]. However, a tightly coupled issue associated with these approaches is packet loss ratio which will increase if neighbor node with minimum reliability factor is selected as a relay node. Therefore, an efficient and reliable mechanism for energy load balancing and communication is needed to be developed to address the aforementioned issues in WSNs without any change in the infrastructure of wireless technologies.

Moreover, the research community suggests various techniques to resolve the load balancing issue of WSNs and prolong the network lifetime. Although, these techniques are efficient at some stage to manage the load balancing issue, but they have some side effects on the network performance such as high communication cost, end-to-end delay, packet lost ratio, throughput, and individual sensor devices lifespan, etc. Therefore, our proposed scheme works based on (1) utilizing the maximum of the shortest path(s); (2) conducting a smart load balancing across multiple paths (particularly those with maximum residual energy, minimum round trip time (RTT) and minimum average packet loss ratio); (3) minimum load on nodes having low residual energy Er and lowest round trip time (RTT) RTTi ; and (4) maximum load on paths with least hop count (Hc value), maximum Er and minimum Nv value; which is largely not considered in many published models, in order to improve the performance and energy efficient for the networks that have limited resources in terms of average packet delivery ratio, networks lifetime.

The proposed model will investigate the potentials of the weighted ensemble based load-balancing scheme in resolving the aforementioned issues such as end- to-end delay; packets deliver ratio, and maximum lifetime. In this paper, a weighted-ensemble based load-balancing scheme is presented for WSNs to address the aforementioned issues. In the proposed scheme, different weight-ages are assigned to various metrics of every member node such as residual energy E_r, Hop-count H_c, crucial N_v (from net- works connectivity perspective) and First Hop Neighbor Round Trip Time ($FNRTT_i$). Moreover, the proposed scheme adopts a sequential approach such that first packet is sent through the most optimal path that is computed using those metrics. Second packet is forwarded on the 2nd optimal path available in the operational WSNs.

The remaining paper is organized as follows. In sub-sequent section, a brief literature review is presented. In Section 3, the proposed mechanism is described in detail whereas results discussion is presented in subsequent section i.e., Section 4. Finally, concluding remarks and future directives are provided.

II. Related Works

In WSNs, which have various constraints, management of load balancing and optimized energy consumption of sensor devices are considered a vital role in enhancing the prolong network lifespan with better end-to-end delay, packet delivery, throughput, and packet loss ratio. Based on that, there is need to develop an efficient communication infrastructure of sensor devices with uniform load balancing to maximize its performance with optimal resource utilization, specifically for WSNs that have limited resources. To address this issue, many schemes have been proposed in the literature. Thus, in this section, the existing techniques of WSNs load balancing are comprehensively overviewed with their merits and demerits.

In the literature, there is an important question, How to develop an efficient load balancing scheme with minimal resources and maximum results for WSNs or IOTs? Proposing a new scheme based on a uniform load or traffic distribution approach over multi-path and multi-hop communication infrastructure to share the collected data of sensor devices in the network is a powerful solution for this issue in networks with limited resources like IOTs and WSNs. In this manner, we used both optimal-path identification scheme and gradient enabled scheme in the directed diffusion method, where server module or base station Sj chooses the optimal operational path; in this case, it will impose on the ordinary or member nodes Ci to utilize this path as a continuous communication link between sensor nodes. This path is used by various devices interested in transmission of data until one or more of these devices consume their on-board power completely or becomes not operational due to other reasons [13], [14]. Generally, this scheme relies on a single communication path, consequently, the lifetime of both WSNs and ordinary devices C_i is very short, that is similar to the shortest path-enabled scheme where packets are forward on this path as long as all nodes reside on it are operational. Likewise, a shortest path-enabled approach was presented to resolve coverage area problem in mobile relay node(s) [15]. A common problem associated with these schemes is the selection and reinforcement of another optimal path that is needed in scenarios where existing path is no longer available or not reliable for further communication. This process is energy starving, time-consuming and costly for the resource-limited networks specifically WSNs and IoTs. A greedy approach enabled algorithm was used to develop a uniform load balancing to spread the traffic uniformly and enhances overall lifetime of the networks [16]. In the same way, Zheng et al., [17,18], proposed a new scheme by computing the transmission capabilities for every sensing device Ci by their neighboring devices based on processing cost, hop count Hc, residual energy Er and round trip time (RTTi). At the beginning, to ensure reliable wireless communication, source device assigns a neighbor device Ci with lowest Hc value over others having highest Hc values. In addition, the scheme proposes that every device Ci should find at least 4 shortest and reliable paths, and forward maximum number of packets on the most two reliable and optimal routes. Therefore, unlike previous shortest path schemes, devices Ci available in these paths do not deplete their on-board battery more rapidly than other schemes through assigning weight-age factors accurately do. Software-Defined Wireless Sensor Network (SDWSN) architecture for load balancing of IoT was proposed by Cui et al. [19]. The proposed technique uses a centralized Software Defined Network (SDN) with flexible traffic management to minimize the energy consumption of participating IoT devices in the network. Moreover, for real traffic monitoring in the network, the authors also used the Open Flow protocol framework to verify the performance of the network. Qin et al. [20] proposed the software-defined network framework for load balancing in WSNs and IoT. Furthermore, they considered the emergency control system in the urban area to verify the feasibility of the proposed scheme.. Based on the results, they found that energy based traffic spreading approach have better performance as compared to other schemes.

Additionally, another study by Yousif et al., [21], they applied a weighted optimal path for loading balancing to solve the problem of uniform traffic distribution in wireless sensor networks. Another scheme known as aware load balancing scheme was proposed [11] to improve the network lifetime with ensuring the least possible percentage of end-to-end delay and packet delivery ratio. The performance of these schemes was exceptionally well, particularly in terms of networks lifetime enhancement metric, but it has some disadvantages like longer path selection to send maximum packets with continuous rate consumes extra energy and this creates network overhead.

The literature of WSNs in terms of load balancing schemes contain various cluster-based schemes or hierarchical networking infrastructures that have been used in the recent past to address the load balancing issues for networks with limited resources such as end to end delay, packet lost ratio, throughput, and sensor devices lifespan, coverage area, etc. [9]. Most of the researchers used the LEACH protocol in their proposed models to address the issue of load balancing in WSNs. However, one of the most difficult challenges in this type of networks is the identification the optimal cluster head node (CH), which is responsible to send the collected data from sensor devices to the concerned base station. Although, LEACH protocol plays significantly will in the formation of the cluster head nodes in the network, but it has some disadvantages like dynamic cluster head nodes selection with continuous rate consumes extra energy and creates network overhead. In the same way, Zhang et al. [22] proposed a new load-balancing scheme for homogeneous networks, where both cluster head nodes (CHs) and member nodes Ci have the same features in terms of processing power and communication. The work of this scheme is based on moving the sensor nodes from CHs that have a heavy load to CHs with lowest load. The movement of sensor node is based on its location. In other word, it depends on the distance of sensor nodes, from the nearest CHs with highest volume of residual energy Er. A load-balancing scheme based on weighted based traffic was developed to solve the networks lifetime issue that have limited resources. To solve this issue, this scheme perform by selecting the CH based on its residual energy and number of deployed nodes in its vicinity. A stochastic distribution based traffic distribution scheme was proposed by Liao et al. [23], which aimed to generate a new model based on a uniform energy consumption throughout the networks. In the same way, another schemes based on tree approach were proposed to solve the load balancing problem and to establish a reliable communication infrastructure [18]. However, the majority of these schemes are either application specific or overlay complex or expensive due to change in the existing communication mechanisms. Therefore, a reliable load balancing mechanism is needed to resolve this issue with available resources.

Theory of information enabled load balancing scheme was presented to distribute traffic uniformly through a proper scheduling of modulation and routing mechanisms [24]. Additionally, Preethiya et al. [25] have proposed mobile double cluster-head enabled particle swarm optimization approach to address uniform traffic distribution challenge in heterogeneous WSNs which is divided into four phases such as CH selection, scheduling of CHs, mobility and handover predication of a specific CH in the operational WSNs environment. However, complexity and implementation in realistic environments are among common issues associated with this approach. Zarin et al. [26] have presented a central controller node-enabled approach to distribute traffic uniformly throughout the networks where radio signals are utilized to balance energy consumption of active nodes in an operational heterogeneous WSN. A common issue with this approach is that reliable communication among mobile nodes and base station is not guaranteed. Similarly, a centralized optimal task allocation scheme is proposed to spread packets uniformly among available paths in the operational networks [27]. To realize this, linear programming and distributed optimal task allocation algorithms were integrated and utilized to minimize energy consumption of sensor nodes.

A. *Problem Statement*

With notable differences of existing studies [11], [7], [25], many schemes have been proposed for the load-balancing problem in WSNs. These schemes have not taken into consideration some important parameters like minimum possible average end-to-end delay and latency or maximum packet delivery ratio in order to ensure a reliable and optimal load-balancing scheme with an ideal wireless communication.

In addition, Existing load balancing schemes were either performed on CH level, particularly in cluster based infrastructures, or in multiple path based scheme usually through Er, as in flat networks, but a generalized load balancing methodology, that is applicable in different networking infrastructures and topologies, is not addressed yet. Therefore, the proposed work will be focused on the development of a reliable neighborhood-enabled load balancing and communication scheme and its implementation in real environment of the resource-limited networks. Furthermore, the proposed scheme should be suitable and implementable in various topological infrastructures of WSNs. Additionally; the proposed work should thoroughly investigate criticality or importance of device C_i particularly from networks connectivity perspective. Criticality of various nodes in the operational WSNs is depicted in Fig. 1 where communication activity of sensor nodes A, B, C and D are primarily dependent on the availability of sensor node H. Therefore, a reliable communication mechanism is needed to be developed which should bound each and every sensor node transmit packet to a non-critical nodes (if available). For example, sensor nodes E, G, F and I should be forced to avoid transmission of packets via sensor node H as alternative routes are available for these nodes in Fig. 1.

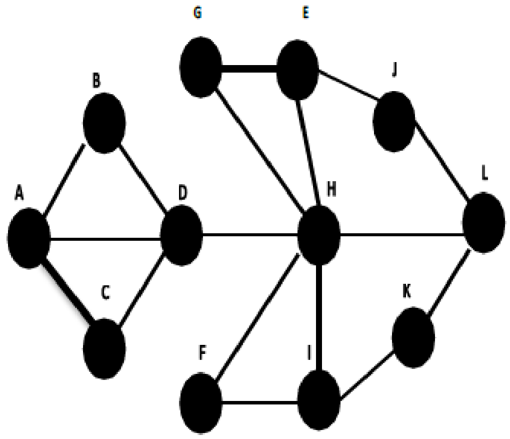


Fig. 1. Importance of Ordinary Sensor Nodes from Networks Connectivity Perspective.

III. Computational Mechanism of Sensor Nodes Criticality Factor (NV)

Importance (criticality factor from the networks connectivity perspective) of a sensor node C_i $WSNs$ is a one-step process that is criticality of a member C_i is computed only once i.e., when the networks deployment process is complete. Generally, the base station module is responsible to compute criticality factors of each and every member node C_i and share it with these nodes. For this purpose, various concept of graph theory (preferably planner graph) are used to compute this factor. Higher value of a sensor node C_i criticality factor N_v is an indication of how important that node is from the WSNs connectivity perspective? For example in Fig. 1, criticality factor value N_v of sensor node H is higher than other nodes as a certain portion of the underlined network (i.e., A, B, C and D) are completely dependent on its active- ness for the transmission of packets. Moreover, this factor play a vital role to distribute traffic in a biased but efficient manner (from residual energy E_r perspective). Therefore, the proposed load balancing mechanism should be smart-enough to minimize traffic flow on paths or neighboring sensor nodes C_i, which have higher values of criticality factor than other nodes in the operational WSNs. A detailed description of the methodology that is used to calculate criticality of sensor nodes $C_i \in WSNs$ is given below.

A. Calculation of Levels in WSNs

The computation process of a sensor node C_i criticality factor is initiated with the calculation of levels in WSNs preferably before and after removal of the concerned node C_i. For this purposed, a well-known technique called logical networks abridgment (LNA) is adopted which is used to generate loop-free description of the underlined WSNs [28]. Moreover, the concept of planner graph is utilized to develop a proper sketch of the deployed WSNs as shown in Fig. 2 where various level are identified such as level-1, level-2, etc. Usually, the deployed WSNs physical infrastructure is represented by Level-1, which is $1^{s}t$ level which is generated through the planner graph concept. An active sensor node *Ci* and communication link between two neighboring nodes in WSNs corresponds to a vertex and an edge in a graph respectively. A closed ring in the underlined graph should be a vertex in the next level as shown in Fig. 2 where an edge in the next level is subjected to neighborhood. A sensor node‘s levels calculation process is assumed to be completed **iff** a graph with no loop or cycle, a closed proximity of three or more nodes, is generated as represented in green at level-3 in Fig. 2. Level-2 graph is generated from level-1 where every vertex belongs to this level corresponds is created only if there is a cycle in level-1 graph and an edge is drawn **iff** there exist a pair of vertices belong to level-1 reside in closed proximity or neighbors. Moreover, these pair of vertices should belong to a similar bi-connected graph. In scenarios where these cycles do not belong to the same bi- connected graph then.

1) An edge in level-2 is drawn if and only if these cycles have a common vertex or node, which is known as cut- node in graph theory.

2) If these cycles do not have any common node(s), then edge should not be added.

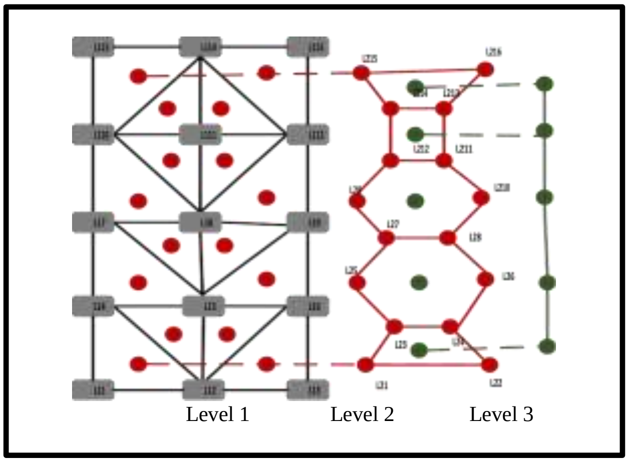


Fig. 2. Sensor Node Criticality Metric Computation Mechanism.

The development process of Level-2 from Level-1 is identical to that of a dual graph where a vertice, which is surrounded by at-least three vertices (preferably to form a ring or cycle), is considered as a vertice in the next level exterior nodes are excluded and interior nodes are considered as shown in Fig. 2. Similarly, Level-3 vertices and edges are formed using this procedure. This process of levels computation is repeated until a loop free structure (a planner graph with no cycles) of the underlined graph is obtained as shown in green in Fig. 2.

After computing number of levels in the entire WSNs, the next phase is calculate these levels for each and every member node *Ci* ∈ *WSNs*. For this purpose, a sensor node for which levels are to be computed is assumed to be removed (temporarily) from the WSNs. For example in Fig. 1, if levels are to be computed for sensor node say 'A', then this node is assumed as removed or non-member from this network which becomes as shown in Fig. 3. Then the above mentioned process for whole networks levels computation is repeated for this network. In this way, levels of whole networks and member nodes are computed.

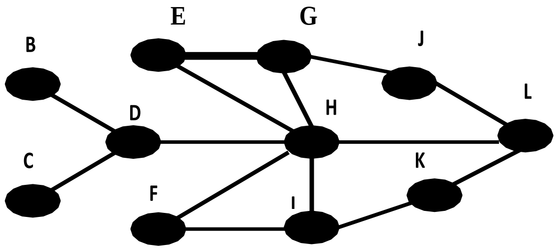


Fig. 3. Level Computation of an Individual Sensor Node.

B. Criticality of Member Sensor Nodes: Calculation Methodology

Criticality metric of a sensor node plays a vital role in prolonging the WSNs connectivity and, hence, enhancing their lifespan. Criticality metric of sensor nodes C_i WSNs is computed using the following formula.

$$v_n = \frac{N_B}{N_A} x \frac{L_{B+C}}{L_{A+C}} \quad (1)$$

Where V_n is the criticality factor of n^{th} sensor node in an operational WSN. N_B and N_A represents number of nodes in network before and after removing n^{th} node respectively. For example, there are twelve nodes in WSNs as shown in Fig. 1. Whereas eleven nodes remains after removing sensor node A from the underlined WSN as shown in Fig. 3. Likewise, L_B and L_A represent levels in WSN before and after removing that node respectively. For example in Fig. 4, both networks have same number of sensor nodes i.e., five. Networks-A is a flat network with single level i.e., each node has a single path of communication. However, in network-B, each sensor nodes has two different paths available for communication. Removal of a node, as described above in level calculation mechanism, from a deployed WSN is used to examine how networks connectivity is effected if that particular node consumes it on-board battery completely. For example, if we remove sensor node C from both networks then nodes belong to network-A will not able to communicate with each other whereas those belong to network-B are able to communicate, as alternative path is available as shown in Fig. 4. Once criticality metric of every sensor node C_i is completed then the base station S_j module shares these values with its member sensor nodes. Thus, each and every sensor node stores criticality values of its neighboring devices along with other information.

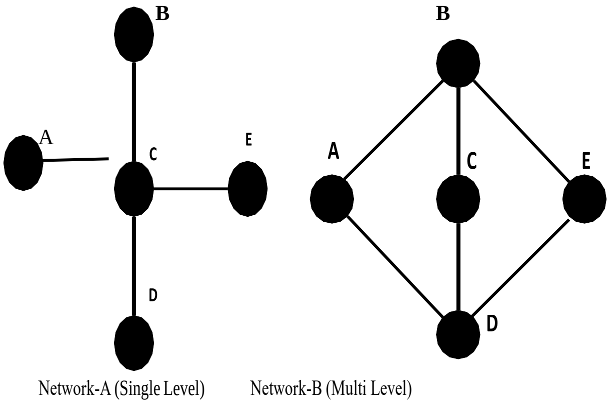


Fig. 4. Effects of Levels on the Networks Connectivity.

IV. Proposed Neighborhood-Enabled Load Balancing Scheme for WSNs

A neighborhood-enabled load-balancing scheme, which is specifically designed for the wireless sensor networks, is presented to distribute traffic uniformly across all available paths in an operational network where every sensor node is bounded to send packet via reliable and non-critical paths or nodes. To achieve this goal, various routing metrics related to the neighboring nodes are used such as hop-count H_c, residual energy E_r, Round Trip Time (RTT) and sensor node criticality N_v.

A. Hop-Count Discover Phase

Hop-count information plays a vital role in resolving the ping-pong (multiple transmission of same message again and again) between neighboring node in operational WSNs environment. For example, a sensor node will discard a message received from neighboring device whose hop-count value is less than its own. In this phase, a control packet, Msg_{hd}, is generated and broad casted by the base station module S_j, with value of $H_c = 0$ as base station is ultimate destination of all packets, which is received by its neighboring nodes. These nodes update their H_c according, that is $H_c = 1$, and broad-cast an updated version of that message Msg_{hd}, i.e., with value of $H_c = 1$, after its back-off time expires. Back-off timer is used to minimize the collision probability of these messages.

$$T_b(C_i) = rand(0-1000) + min(H_c(C_i), \delta) \quad (2)$$

The idea of adding an *Hc* value or δ with the generated random number is to further minimize (or avoid if possible) the packets collision probability of neighboring nodes as usually, these nodes have different *Hc* values. However, in scenarios where back-off timer Tb of two or more neighboring nodes are similar, then these nodes *Ci* should re-compute their back-off timer Tb accordingly. Moreover, parameter δ is an application dependent value whose values are different for different topological infrastructures such as in flat networks its value is in between 5 and 15 whereas between 2 and 5 for the hierarchical WSNs. This process is repeated by each and every member sensor node Ci WSNs until *Hc* value is collected by every member device in the WSNs.

Moreover, every sensor node C_i keeps a record of it neighboring nodes along with their hop-count values. For this purpose, a simple table is maintained where this information is recorded from time to time. Separate session is not needed for this purposed rather it is performed as sub part of hop- count discovery phase. For example, when a sensor node receive message Msg_{hd} that contains information about it source such as hop-count and MAC-ID. Every sensor node records this information in a table, which is used to identify optimal and reliable neighboring nodes for the transmission of packets.

B. Finding Optimal Neighboring Nodes

In the proposed setup, an optimal neighboring node C_i *WSNs* is defined as a sensor node with maximum residual energy E_r and minimum possible (hop-count H_c, criticality N_v & Round Trip Time (RTT) values. In next phase, these nodes find two optimal neighbors or paths, preferably with minimum hop-count H_c, lowest RTT_i value, minimum N_v value and maximum residual energy E_r. These paths are used in the biased way (such as forwarding maximum traffic on most optimal path i.e., 70% whereas 30% on other path) to increase average packet delivery ratio and decrease the overall packet lost ratio and end-to-end delay of packets in an operational WSN. It is to be noted that packet delivery ratio is improve if majority of the packets are sent via most optimal path. Moreover, it is quite likely that packet loss ratio is minimize due to forwarding packets over a most reliable path or neighbor.

In order to identify optimal neighboring node(s), different weight-ages are assigned to the hop-count H_c, residual energy E_r, RTT and Criticality N_v values. Initially, maximum weight-age is assigned to the hop-count and criticality N_v metrics of various neighboring nodes to forward maximum packets or traffic through the shortest and less critical path. Moreover, specifically at this stage, residual energy E_r of all neighboring nodes are similar, therefore, it will not affect the optimal neighbor's' selection process irrespective of higher or lower weight-age value assigned to it. Optimal neighboring node selection process is subjected to the following (3):

$$Opt_{Neighbor}=Min(W_1\times H_c+W_2\times N_v+W_3\times E_r+W_4\times RTT_i) \qquad (3)$$

Whereas $W1$, $W2$, $W3$, $W4$ are the weight-age factor assigned to the aforementioned metrics. In proposed setup, initially these values were assigned i.e., $W1$ = 30%, $W2$ = 40%, $W3$ = 10% and $W4$ = 20%. N_v value represents the criticality of a member node from the resource limited networks connectivity factor i.e., how important is a particular node C_i as far as WSN‘s long term connectivity is concerned. As described above that a higher value of criticality N_v associated with a node is an indication that a certain portion of the underlined WSN either partially or completely rely on this node i.e., if it becomes non-active (by consuming its on- board battery) then the portion of WSN will be disconnected or will not be able to communicate properly with the base station module S_j. Moreover, $Opt_{Neighbor}$ is the node C_i with (probably) a shortest path to the destination module i.e., base station.

Initially, packets generated by various source nodes C_i are distributed uniformly between two optimal neighboring nodes i.e., 50% packets forwarded through $Opt_{Neighbor}1$ and $Opt_{Neighbor}2$ respectively. However, the proposed scheme should avoid (if possible) critical nodes even though these are the most optimal neighbors. A neighboring sensor node C_i is assumed to be critical **iff**

1) Certain portion(s) of the underlined WSNs are either partially or completely dependent on that neighbor for the transmission of packets.

2) It acts as the shared optimal path among various nodes in the WSNs.

3) It has consumed approximately 80% of its on-board battery.

Moreover, if two or more neighboring nodes have a similar optimal value, then the selection process is random. Forwarding packets on most reliable path, a path with zero critical node, results in enhancing lifespan of the underlined constraint oriented networks. Additionally, the WSNs become connected for maximum possible duration without changing in the technological infrastructure.

The weight-age ratio presented in equation. 3 is maintained by every sensor node C_i until one or more of its neighboring nodes consume 75% of their on-board batteries. 75% threshold value is selected due to the fact that this node will be active for maximum possible duration. This node is now dedicated to forward data or packets of those nodes that have no other path(s) or neighbor(s) toward the base station module S_j. As soon as a sensor node C_i broadcast a message which indicates it critical condition i.e., that it has consumed 75% of the available power, then neighboring node(s) re-adjust their packets forwarding schedule or methodology accordingly. In this scenario, criteria for finding optimal neighboring nodes is revised in equation 3 by adjusting their weight-age factors such as residual energy E_r is assigned highest weight-age followed by nodes criticality value N_v. These weight-ages are adjusted as follows. $W1$ = 10%, $W2$ = 30%, $W3$ = 40% and $W3$ = 20%. Now, maximum packets are routed through those neighbors, which have maximum possible residual energy E_r and minimum criticality value N_v. This re-adjustment policy not only resolves end- to-end delay issue that is linked to the WSNs but at the same time improve their lifespan considerably.

C. Neighborhood-Enabled Load Balancing Algorithm

Generally, criticality of a sensor node C_i is high if it has maximum number of neighboring nodes particularly those nodes that have the capacity to communicate directly with it. Thus, it is quite likely that criticality of node C_i *WSNs* is strongly correlated to its congested neighborhood. A node with maximum neighbors has a higher probability to forward a slightly higher ratio of packets than other nodes, i.e., node(s) with minimum neighboring nodes. Therefore, critical node(s) consume their on-board battery more rapidly than other nodes due to the heavy load. Therefore, a trade-off metric is implemented to relieve these nodes either partially or completely (if possible) with an acceptable performance degradation ratio in terms of transmission delay, energy consumption etc. The proposed load-balancing algorithm, which is presented below as Algorithm. 1, bounds every sensing device to thoroughly evaluate its neighborhood in terms of criticality, reliability, power capacity and distance before initiating a communication session.

D. Computational Complexity of the Proposed Load Balancing Scheme

Computational complexity is an evaluating metric that is used to describe the potentials and requirements of an underlined algorithm, in terms of time and space, to resolve a specific problem. Complexity of an algorithm has a direct co-relation to its time and space requirements.

V. Results of Simulation and Discussion

In this section, a detailed analysis of the simulation results is presented to evaluate the performance of the proposed neighborhood-enabled load balancing algorithm in terms of end-to-end delay, packet delivery-ratio, uniform load balancing (where possible), energy efficiency and residual energy (*Er*). The proposed LBS is compared with the field proven techniques i.e., the shortest path algorithm, energy based traffic spreading approaches [8], [29], opportunistic routing [30], and vulnerability aware routing [11], [12]. These algorithms were implemented in OMNET++ [31], an open source simulation environment that is specifically designed for the resource-limited networks. It is to be noted that similar processing and communication power, on-board batteries, a single gateway module were used both for the proposed and existing schemes. Moreover, these schemes were tested using different topological structures such as graph based topology, tree based, random-top, and random- center. Additionally, channels or paths delay factors; both path loss ratio and interference are assumed to be constant for the WSNs. A WSN whose nodes (preferably n in this case) are distributed randomly in a deployment area of 500m * 500m as shown in Table I. Moreover, we have assumed that a sensor node either communicates directly with the base station module **iff** it resides in closed proximity of the base station or through various relay nodes using multi-hop communication. Additionally, for simplicity the capacities of on-board battery and transceivers are similar to the wasp- mote agriculture pro-board capacity [32] i.e., 52000 mAh and 400 meters respectively. To streamline the proposed neighborhood-enabled load balancing scheme applicability in the realistic environments of WSNs, well-known ratios are used both for packets transmission and reception i.e., P_{TX} = 91.4mW and P_{RX} = 59.1mW respectively. More- over, initial hop-count of ordinary nodes C_i is equal to ( ). In the proposed scheme, every node C_i is needed to inform its neighbor's **iff** its residual energy E_r is less than the defined threshold value i.e., 85% in this case. In proposed setup, the distance between various nodes C_i is kept around 400 meter. Moreover, criticality factor of WSNs and individual sensor nodes is computed only once i.e., precisely after the deployment process. As described above, this process is performed by the gateway or base station module.

The proposed scheme were thoroughly examined against various fields proven schemes (preferably load balancing and routing) in term of residual energy E_r, lifetime, end-to-end delay, average packet delivery ratio and sent or forwarded packets.

A. Uniform Energy Consumption

The proposed scheme was designed and developed not only to distribute traffic uniformly across the networks (WSNs), specifically via reliable communication path(s), but at the same time generated traffic is distributed in such a way that each and every node consume approximately similar power i.e., their residual energy is approximately similar (preferably of neighboring nodes). The proposed scheme performance is far better than existing schemes as shown in Fig. 5.

TABLE I. Homogeneous WSNs Simulation Parameters Setup

Parameters	**Values**
WSN Deployment Area	500m * 500m
Sensor Node	40, 200, 300, 500
Base Station	One
Initial Energy (E_s)	52000 mAh
Residual Energy (E_r)	E_s-$E_{consumed}$
Packet Transmission Power Consumption (P_{TX})	91.4 mW
Channel Delay (Ch_{delay})	10 millisecond
Packet Receiving Power Consumption (P_{RX})	59.1 mW
Idle Mode Power Consumption	1.27 mW
Sleep Mode Power Consumption	15.4 μW
Transceiver Energy (T_i)	1 mW
Transmission Range (T_r)	200m
Receiving Power Threshold (RTS_n)	1024 bits
Packet Size (P_{size})	127 bytes
Hop Count (H_c) of Base Station	0
Initial Hop Count (H_c) of Sensor Nodes	
Initial Criticality Value (V_n) of Sensor Nodes	0
Maximum Distance between Nodes	100m
Sampling Rate of sensor nodes	30 Seconds

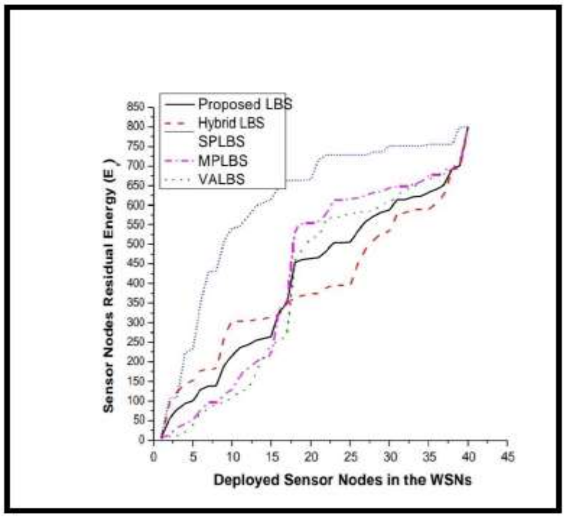


Fig. 5. Residual Energy of Sensor Nodes belong to the Operational WSNs.

Algorithm 1 Proposed Neighborhood-enabled Load Balancing Algorithm

```
Require: Forwarding of Data that is Collected by
    Sensor Nodes
Ensure: Return Optimal Neighbor or Path
$E_r$← 100%
$V_n$←∞
$H_c$ ← 0
$RTT_i$    0
Local Function ()
    Packet ← Either Generated or Forwarded
    if($E_r$ > 25%) then
$P_{opt}$ = OptimalNeighbor - 1(numberofneighbors)
    else
$P_{opt}$ = OptimalNeighbor - 2(numberofneighbors)
    Packet ← ($P_{opt}$)
  end LocalFunction
OptimalNeighbor-1 (number)
    NCvalue ←0
  $N_{opt}$←∞
    $w1$ ← 30%
    $w2$ ← 40%
    $w3$ ← 10%
    $w3$ ← 20%
    for I ← 0 to number do
$C_{cur}$ ← $w1(H_c)_i$+$w2(N_v)_i$+$w3(E_r)_i$+$w4(RTT)_i$
      if ($C_{cur}$ < $N_{opt}$) then
        $Nopt$ ← $Ccur$
      elseif ($C_{cur}$ = $N_{opt}$) then
        $N_{opt}$ ← rand($N_{opt}$, $C_{cur}$)
      end if
    end for
      return $N_{opt}$
END NEIGHBORDISCOVERY-1
OptimalNeighbor-2 (number)
    NCvalue ←0
  $N_{opt}$←∞
    $w1$ ← 10%
    $w2$ ← 30%
    $w3$ ← 40%
    $w3$ ← 20%
    for I ← 0 to number do
      $C_{cur}$ ← $w1(H_c)_i$+$w2(N_v)_i$+$w3(E_r)_i$+$w4(RTT)_i$
      if ($C_{cur}$ < $N_{opt}$) then
        $Nopt$ ← $Ccur$
      elseif ($C_{cur}$ = $N_{opt}$) then
        $N_{opt}$ ← rand($N_{opt}$, $C_{cur}$)
    end for
      return $N_{opt}$
END NEIGHBORDISCOVERY-2
end
```

B. WSNs Lifetime

A prolonged lifetime of the resource-limited networks, WSNs and IoTs, is a challenging issue up to date. In literature, various mechanisms have been proposed to address this issue. It is evident from Fig. 6 that the proposed scheme is ideal solution for WSNs as it outperform the existing schemes. The WSNs lifetime is enhanced by the proposed scheme that is subjected to strong emphasis of the proposed neighborhood-enabled load balancing scheme on the avoidance (if possible and applicable only) of critical nodes during the communication activity. Moreover, as WSNs lifetime is subjected to the lifespan of individual nodes. Therefore, a brief analysis of an individual node lifetime is presented in Fig. 7. Lifetime of both individual node and whole WSNs is inversely proportional to the sampling rate i.e., a highest sampling interval leads to the minimum lifespan of WSNs.

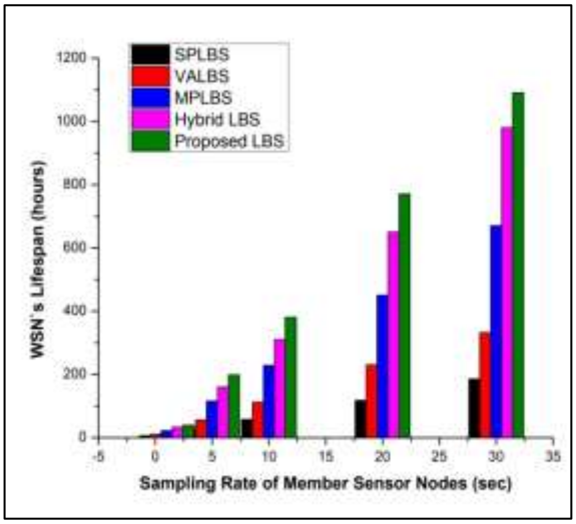


Fig. 6. Lifetime Comparison of the Proposed Neighborhood-Enaod-Enabled LBS and Field Proven Algorithms.

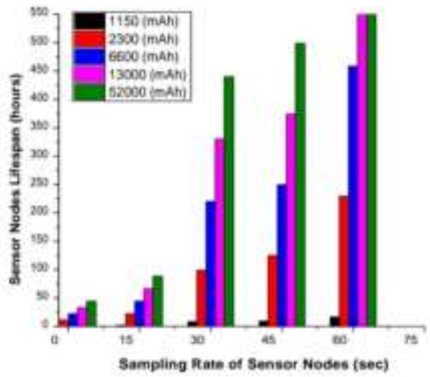


Fig. 7. Sensor Nodes Capacities on different Standard of Waspmote On-board Batteries i.e., 1150, 2300, 6000, 13000, and 52000 mAh.

C. End-To-End Delay

End-to-End delay is a basic measure to evaluate the performance of a load-balancing scheme specifically in the resource limited networks environments. A comparison of the pro- posed neighborhood-enabled LBS and existing algorithms in terms of average end-to-end delay performance metric is presented in Fig. 8, which shows that the proposed scheme has the lowest possible end-to-end delay metric than existing schemes specifically on various WSNs densities that number of member nodes. The proposed LBS achieves this milestone by selecting an appropriate path for an individual packet by considering different parameters of the neighboring nodes such as criticality factor, residual energy, and hop count etc. Additionally, Fig. 8 shows that end-to-end delay factor is directly proportional to the networks size, channel delay, queuing delay, transmission delay, receiving delay, and the location of source and destination devices.

D. Average Packet Delivery Ratio

Average packet delivery ratio is another important performance metric that is utilized to evaluate various load balancing schemes in the realistic environment of WSNs. Fig. 9 shows the comparison of average packet delivery ratio of the proposed neighborhood-enabled LBS with the existing schemes, which shows that the proposed scheme has maximum possible average packet delivery ratio. Moreover, these results were verified on various WSNs infrastructures such as different number of nodes C_i and topologies. It is to be noted that the proposed neighborhood-enabled LBS algorithm achieves the highest ratio by utilizing a smart neighbor (preferably relay node) selection criteria and thus reducing the packet loss ratio.

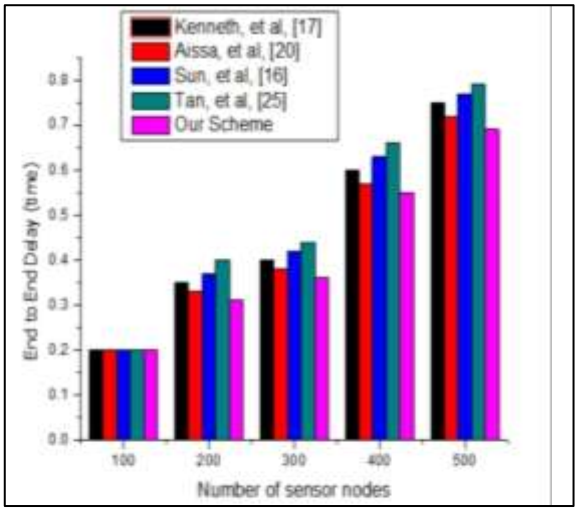


Fig. 8. End-to-End Delay Analysis Statistics of the Proposed and Existing Load Balancing Schemes.

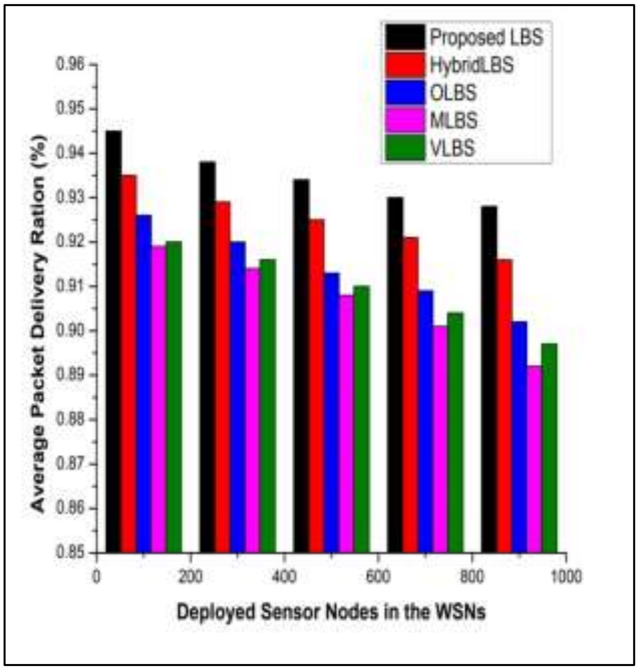


Fig. 9. Average Packet Delivery Ratio Analysis Statistics of the Proposed and Existing Load Balancing Schemes.

E. Throughput Analysis

An alternative evaluation metric is the throughput that is defined as total number of successfully delivered packets to the base station module in the WSNs. A load-balancing algorithm is considered as efficient and reliable **iff** it has achieved maximum throughput with available resources. The proposed scheme performs better than existing scheme in different realistic scenarios of the WSNs as shown in Fig. 10. The proposed neighborhood-enabled LBS has achieved the milestone of maximum possible throughput by bounding every node C_i to send packet(s) via most reliable and preferably shortest path(s) or neighbors.

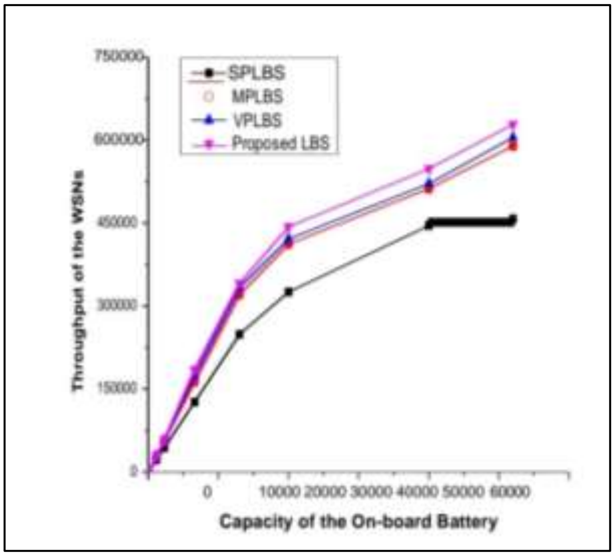


Fig. 10. Throughput Analysis Statistics of theProposed and Existing Load Balancing Schemes.

VI. Conclusions

In this paper, a neighborhood-enabled load-balancing scheme was presented for wireless sensor networks to resolve various issues associated with the existing schemes. Moreover, the proposed scheme did not compromise on other performance metrics such as end-to-end delay, lifetime, average packet delivery ratio and throughput of the concerned WSNs. The proposed neighborhood-enabled load balancing scheme assigns different weight-ages to various metrics or parameter of every neighboring node such as residual energy E_r, Hop-count H_c, crucial N_v (from networks connectivity perspective) and first hop neighbor round trip time $(FNRTT_i)$. Moreover, the proposed load-balancing scheme adopts a sequential approach such that first packet is sent through the most optimal path that is computed using those metrics and next packet is forwarded on the 2^{nd} optimal path available in the operational WSNs. Simulation results have verified the exceptional performance of the proposed load balancing scheme against the existing LBS approaches particularly in terms of end-to-end delay, average packet delivery ratio, lifetime (both individual and whole WSNs) and throughput.

References

[1] X. Liu, R. Zhu, A. Anjum, J. Wang, H. Zhang, and M. Ma, “Intelligent data fusion algorithm based on hybrid delay-aware adaptive clustering in wireless sensor networks,” Future Generation Computer Systems, vol. 104, pp. 1–14, 2020.

[2] Almaiah MA, Dawahdeh Z, Almomani O, Alsaaidah A, Al-khasawneh A, Khawatreh S. A new hybrid text encryption approach over mobile ad hoc network. International Journal of Electrical and Computer Engineering (IJECE). 2020 Dec;10(6):6461-71.

[3] F. Wu, X. Li, L. Xu, P. Vijayakumar, and N. Kumar, “A novel three-factor authentication protocol for wireless sensor networks with iot notion,” IEEE Systems Journal, 2020.

[4] M. Sajwan, D. Gosain, and A. K. Sharma, “Camp: cluster aided multi-path routing protocol for wireless sensor networks,” Wireless Networks, vol. 25, no. 5, pp. 2603–2620, 2019.

[5] X. Liu, T. Qiu, and T. Wang, “Load-balanced data dissemination for wireless sensor networks: A nature-inspired approach,” IEEE Internet of Things Journal, vol. 6, no. 6, pp. 9256–9265, 2019.

[6] L. Cheng, J. Niu, C. Luo, L. Shu, L. Kong, Z. Zhao, and Y. Gu, “Towards minimum-delay and energy-efficient flooding in low-duty-cycle wireless sensor networks,” Computer Networks, vol. 134, pp. 66–77, 2018.

[7] S. K. Singh, M. Singh, D. K. Singh et al., “Routing protocols in wireless sensor networks–a survey,” International Journal of Computer Science & Engineering Survey (IJCSES), vol. 1, no. 2, pp. 63–83, 2010.

[8] A. Laouid, A. Dahmani, A. Bounceur, R. Euler, F. Lalem, and A. Tari, “A distributed multi-path routing algorithm to balance energy consumption in wireless sensor networks,” Ad Hoc Networks, vol. 64, pp. 53–64, 2017.

[9] H.-Y. Kim, “An energy-efficient load balancing scheme to extend lifetime in wireless sensor networks,” Cluster computing, vol. 19, no. 1, pp. 279– 283, 2016.

[10] A. R. M. Kamal and M. A. Hamid, “Supervisory routing control for dynamic load balancing in low data rate wireless sensor networks,” Wireless Networks, vol. 23, no. 4, pp. 1085–1099, 2017.

[11] A. Sari and E. Caglar, “Load balancing algorithms and protocols to enhance quality of service and performance in data of wsn,” in Security and Resilience in Intelligent Data-Centric Systems and Communication Networks. Elsevier, 2018, pp. 143–178.

[12] R. Khan, “An efficient load balancing and performance optimization scheme for constraint oriented networks,” Simulation Modelling Practice and Theory, vol. 96, p. 101930, 2019.

[13] R. Khan, M. Zakarya, Z. Tan, M. Usman, M. A. Jan, and M. Khan, “Pfars: Enhancing throughput and lifetime of heterogeneous wsns through power- aware fusion, aggregation, and routing scheme,” International Journal of Communication Systems, vol. 32, no. 18, p. e4144, 2019.

[14] Y. Yuan, W. Liu, T. Wang, Q. Deng, A. Liu, and H. Song, “Compressive sensing-based clustering joint annular routing data gathering scheme for wireless sensor networks,” IEEE Access, vol. 7, pp. 114 639–114 658, 2019.

[15] D. Cheng, Y. Song, Y. Mao, and X. Wang, “Lddp: A location-based directed diffusion routing protocol for smart home sensor network,” in Systems and Informatics (ICSAI), 2014 2nd International Conference on. IEEE, 2014, pp. 510–514.

[16] C. N. Abhilash, S. H. Manjula, R. Tanuja, and K. Venugopal, “T shortest path discovery for area coverage (spdac) using prediction-based clustering in wsn,” in In Advances in Artificial Intelligence and Data Engineering. Springer, 2019, pp. 1345–1357.

[17] N. Kim, J. Heo, H. S. Kim, and W. H. Kwon, “Reconfiguration of clusterheads for load balancing in wireless sensor networks,” Computer Communications, vol. 31, no. 1, pp. 153–159, 2008.

[18] B. Touray, J. Shim, and P. Johnson, “Biased random algorithm for load balancing in wireless sensor networks (bralb),” in Power Electronics and Motion Control Conference (EPE/PEMC), 2012 15th International. IEEE, 2012, pp. LS4e–1.

[19] P. M. Daflapurkar, M. Gandhi, and B. Patil, “Tree based distributed clustering routing scheme for energy efficiency in wireless sensor networks,” in 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI). IEEE, 2017, pp. 2450–2456.

[20] R. C. Shah and J. M. Rabaey, “Energy aware routing for low energy ad hoc sensor networks,” in Wireless Communications and Networking Conference, 2002. WCNC2002. 2002 IEEE, vol. 1. IEEE, 2002, pp. 350–355.

[21] C. Schurgers and M. B. Srivastava, “Energy efficient routing in wireless sensor networks,” in Military communications conference, 2001. MILCOM 2001. Communications for network-centric operations: Creating the information force. IEEE, vol. 1. IEEE, 2001, pp. 357–361.

[22] Y. K. Yousif, R. Badlishah, N. Yaakob, and A. Amir, “An energy efficient and load balancing clustering scheme for wireless sensor network (wsn) based on distributed approach,” in Journal of Physics: Conference Series, vol. 1019, no. 1. IOP Publishing, 2018, p. 012007.

[23] H. Zhang, L. Li, X.-f. Yan, and X. Li, “A load-balancing clustering algorithm of wsn for data gathering,” in Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on. IEEE, 2011, pp. 915–918.

[24] Y. Liao, H. Qi, and W. Li, “Load-balanced clustering algorithm with distributed self-organization for wireless sensor networks,” IEEE sensors journal, vol. 13, no. 5, pp. 1498–1506, 2013.

[25] Z. Liu, J. Zhang, Y. Li, L. Bai, and Y. Ji, “Joint jobs scheduling and light- path provisioning in fog computing micro datacenter networks,” Journal of Optical Communications and Networking, vol. 10, no. 7, pp. B152–B163, 2018.

[26] T. Preethiya, A. Muthukumar, and S. Durairaj, “Double cluster head heterogeneous clustering for optimization in hybrid wireless sensor network,” Wireless Personal Communications, vol. 110, no. 4, pp. 1751–1768, 2020.

[27] N. Zarin and A. Agarwal, “A centralized approach for load balancing in heterogeneous wireless access network,” in 2018 IEEE Canadian Confer- ence on Electrical & Computer Engineering (CCECE). IEEE, 2018, pp. 1–5.

[28] W. Yu, Y. Huang, and A. Garcia-Ortiz, “Optimal task allocation algorithms for energy constrained multihop wireless networks,” IEEE Sensors Jour- nal, vol. 19, no. 17, pp. 7744–7754, 2019.

[29] T. Arvanitis, C. Constantinou, A. Stepanenko, Y. Sun, B. Liu, and K. Baughan, “Network visualisation and analysis tool based on logical network abridgment,” in Military Communications Conference, 2005. MILCOM 2005. IEEE. IEEE, 2005, pp. 106–112.

[30] K.-V. Nguyen, P. Le Nguyen, Q. H. Vu, and T. Van Do, "An energy efficient and load balanced distributed routing scheme for wireless sensor networks with holes," Journal of Systems and Software, vol. 123, pp. 92–105, 2017.

[31] X. Zhang, L. Tao, F. Yan, and D. K. Sung, "Shortest-latency opportunistic routing in asynchronous wireless sensor networks with independent duty- cycling," IEEE Transactions on Mobile Computing, 2019.

[32] A. Varga, "Omnet++ discrete event simulato," https://omnetpp.org/, 2018.

[33] Libelium "Wasp-mote agriculture boards," http://www.libelium.com/ products/waspmote/, 2018.

[34] Almaiah A, Almomani O. An investigation of digital forensics for Shamoon attack behaviour in fog computing and threat intelligence for incident response. Journal of Theoretical and Applied Information Technology. 2020 Apr 15;98(07).

[35] Almaiah A, Almomani O. An investigator of digital forensics frequencies particle swarm optimization for detection and classification of APT attack in fog computing environment (IDF-FPSO). Journal of Theoretical and Applied Information Technology. 2020 Apr 15;98(07).

Evaluation of Blockchain-based Data Sharing Acceptance among Intelligence Community

Wan Nurhidayat Wan Muhamad[1], Noor Afiza Mat Razali[2], Muslihah Wook[3], Khairul Khalil Ishak[4]
Norulzahrah Mohd Zainudin[5], Nor Asiakin Hasbullah[6,] Suzaimah Ramli[7]
National Defence University of Malaysia, Sungai Besi, Kuala Lumpur, Malaysia[1,2, 3, 5, 6, 7]
Management and Science University, Shah Alam, Selangor, Malaysia[4]

***Abstract*—Intelligence data are among the critical elements used as a reference for risk-assessment and decision-making regarding national security. The intelligence data are shared among intelligence agencies in the intelligence community in improving the efficiency of their services. Centralised data with central authority is highly exposed to being an easy target of attackers. Leaked or unauthorised access of the intelligence data to a non-intelligence community will bring severe effect to a state. Blockchain as immutable and high-security technology is capable of providing cryptographic data in a decentralised environment and potentially can be applied for data sharing among the intelligence community. However, the acceptance and readiness of users on blockchain usage in the intelligence community are yet to be systematically studied. Considering the statement, this paper proposed an evaluation method to study the acceptance of blockchain technology by integrating a reliable acceptance model for blockchain technology implementation in the intelligence community. The acceptance model consisted of constructs from the Technology Acceptance Model 3 (TAM 3) and Technology Readiness Index 2 (TRI 2) and was analysed using partial least squares structural equation modelling (PLS-SEM). In this study, the result indicates that TAM 3 and TRI 2.0 integration model could contribute to determining the acceptance level in developing blockchain-based intelligence data sharing for the intelligence community.**

Keywords—Technology acceptance model; technology readiness index; blockchain acceptance; PLS-SEM; data sharing

I. Introduction

Information sharing in a community becomes easier with the assistance of technology. The intelligence community also benefitted from this technology advancement by shifting its technique of gathering data from traditional Human Intelligence (HUMINT) to a more sophisticated and advanced method of Signal Intelligence (SIGINT) [1]. Information or data collected at the intelligence centre are varied and could be derived from devices and sensors. Analysed data are essential in providing tactical and operational data to organisation, governments, agencies and warfighters [2], [3]. Intelligence community need efficient information sharing among agencies involved to avoid intelligence failure. Example of intelligence failure is such as missing or inadequate data [4]. The process of intelligence data dissemination in the intelligence community is undeniably complicated and challenging. Ensuring accurate and precise data are appropriately disseminated is essential. There is no doubt that the risk of handling such clandestine and important intelligence data is excessive. Leaked or breached intelligence data could deadly affect country sovereignty which also significantly affect the civilian community such as politics, cultural, economy or even lives [1], [5].

Besides, such data shall only be handled by respected authorised agencies that are recognised as the intelligence community. Unauthorised access of data by non-intelligence agency posits grave effect to not only the intelligence community but also the security of a country [6]. Hence, past studies suggested the implementation of access control to heighten data security. As an example, multi-factor authentication technique [1]. However, in this pervasive usage and advancement of the Internet, such authentication technique is insufficient [7], [8]. Thus, there is a suggestion by researchers to consider blockchain as the additional weapon in preserving better security of data [9].

Past and current studies show a significant result of success through the implementation of blockchain in preserving better security of data [10]. Considering that intelligence data are needed to be shared among agencies in a time-wise manner without neglecting the accuracy, blockchain is good to consider to be implemented. However, the study on the implementation of blockchain in the intelligence community is yet to be done. In addition, doubt in using and accepting new technology remains the biggest challenge in introducing new technology.

Therefore, to overcome such challenges, it is highly recommended to investigate users' readiness and acceptance level towards the usage of new technology [11]. Thus, this paper will propose a conceptual model of acceptance and readiness on blockchain-based access for the intelligence community based on constructs in Technology Acceptance Model 3 (TAM 3) and Technology Readiness Level 2 (TRI 2).

The next section of this article will review the background and relevant literature on the intelligence community, blockchain technology, and acceptance studies includes TAM 3, TRI 2.0, and previous acceptance studies on blockchain technology. The conceptual model and hypotheses development are discussed in the proposed model to evaluate blockchain-based data sharing among the intelligence community. The methodology used, result and findings, discussions and conclusion are presented in the latter part of this article.

II. Intelligence Community

National security is about a state of being free from any external or internal danger or threat to its core values. For example, social threats may include animosity from the nation that share the same border, attack by a radical group, and situation of global economic trends that may affect the welfare of the country. In similar scenario, threats or dangers could be defined as a natural disaster or a viral disease outbreak. Such threats could risk the harmony and sovereignty of the affected country. The government must be ready to mobilise its national security system when a nation faces direct or indirect threats. This is where the intelligence community played its vital role. Intelligence community must capable in providing the information needed by a country for security purposes. The primary role of the intelligence community includes to acquire and perform data analysis and share it with their client such as National Security Council (NSC), Defence Agency and more [4]. Such responsibility is given to them due to the confidential level of the information obtained in ensuring the security of a country. Information acquired are stored as intelligence data and given to any organisation that required it. The literature stated that the intelligence community operated based on the intelligence cycle. The cycle consists of planning and directions, collection, process, analyses production and dissemination [4]. The intelligence community could be any government agencies and organisation involved in the management of intelligence environment for the benefit of the country. Besides, the private sector also plays a crucial role in handling intelligence-related projects or systems with intelligence agencies [1].

III. Blockchain Technology

Big data is an enormous and vast pooled data that is too huge for a conventional database to manage. Data is now a key asset for an organisation. Examples of pooled data are such as climate information, GPS signal, online shopping records and more [12]. With big data, there is a new challenge that arises related to the privacy and security of data [13]. Data should exhibit the CIA attributes, which are confidentiality, integrity, and accessibility to be trusted. However, systems that managed big data are prone to exploitation and have a risk to be compromised [14]. Such risk exposure bound to happen due to the wrong configuration of access control and authentication [15], [16]. The statement shows good configuration access control and authentication is indispensable in preserving data security. This is where blockchain integration in big data management come to the surface. The prior study suggests integrating blockchain in handling data, especially risky and confidential data due to its capability of protecting data [7], [9], [17].

Blockchain is defined as a number of blocks that holding information about the respective chain of the individual transaction where each block is linked to the previous block [18]**.** The linkage of blocks is based on the hash value of the previous block, or also known as the parent block. To illustrate, a block can transverse through the whole blockchain and find back each transaction that has been made through its parent block. Block that first to be created and have no parent is called the genesis [19]. According to [20], blockchain is different from any existing scalable database due to its two main features of, i) cryptography by design and ii) lack of control party. Cryptography by design referred to cryptography implementation in preserving the user identity, ensuring the ledger's integrity and the authenticity of data. The cryptography of each block is differ depending on protocol [19]. The hashing algorithms are implemented as a way to ensure blocks are well-formed, to preserve the security of block being tamper-free and be virtually unbreakable [19].

From a software architecture perspective, blockchain enables the development of a new distributed and decentralised software architecture, where confidential transaction or agreement can be made across the chain with untrusted people [10], [13], [21]. Blockchain's criteria of no-human intervening during a process of transaction made it widely applied in various field. As an example, in public services [22]–[24], healthcare [25]–[27], IoT [12], [28] rather than only on the financial system. Nowadays, usage of blockchain is increasing as its source is made as an open-source, which mean anyone can use the entire history of it or modify it legally without need of paying for the service [29].

Blockchain is proposed as the technology that could give the assurance for the intelligence data integrity since there is no central authority and fully automated that enables a safe manner of information passing. No central authority meaning anyone has rightfully approved the transaction being made. However, blockchain is still not widely used and implemented. Hence lack of awareness among the target group requires a preliminary assessment be done on blockchain acceptance and readiness that need to be addressed before the implementation decision could be made.

IV. TAM 3 and TRI 2.0

A. *Technology Acceptance Model 3 (TAM 3)*

Technology acceptance defined as the willingness of an individual to embrace the usage of new technology as per its designated task [30]. As a result, the Technology Acceptance Model (TAM) is established to investigate an individual's level of acceptance in adopting new technology [31], [32]. To emphasise, established TAM by [33], [34] is widely used as the research model in studies of the determinants of technology acceptance in predicting user acceptance and intentions of embracing new technology from individual's perspective. The determinants of TAM are comprising of Perceived if Usefulness (PU) and Perceived Ease of Use (PEoU). PU is defined as the degree of an individual believes that usage of the respected technology would enhance the job quality and their life. While PEoU focused on the degree of individual believes that usage of specific technology will be less of effort and easy to figure.

Researchers have proven that PU and PEoU have positively affected the attitude of an individual's towards intention to use and acceptance of the technology. Investigation on user acceptance towards the usage of technology has been done for over two decades with several models that have been established. As an example, TAM, the extension version of TAM (TAM 2) and TAM 3 [35]. Researchers suggested the application of TAM 3 due to its ability in investigating new relationship compared to TAM and TAM 2 [35], [36]. TAM 3 posits constructs on measuring individuals' acceptance and adoption of the use of technology which give more illustration on the individual's perspective upon the technology acceptance and exhibit a complete representation of constructs to observe individuals' IT adoption and use [35] thus, also suitable to study the individual acceptance of blockchain.

[35]. The relationships are i) relationship between PU and PEoU, (ii) relationship between computer anxiety and PEoU, and (iii) PEoU with behavioural intentions.

TAM 3 is established theoretically based on four factors of acceptance which is i) Social Influences, ii) Individual Differences, iii) System Characteristic and iv) Facilitating Conditions. These four factors are differently wielded influences towards PU and PEoU determinants [37]. Fig. 1 shows the essential four main criteria of TAM theoretically

All the factors and determinants are clustered into their respective criteria. This is to avoid the cross-influence by PU and PEoU. Social Influences described as representing the importance of people believe in the benefit of system usage. While System Characteristics illustrated by the cognitive instrumental process which people believe in positive advantages acquired upon technology usage. Individual Differences heightened the general belief of individual towards computer and computer usage. The last criteria, of Facilitating Condition represent the perception of external control determinants related to the availability of support and resources of an organisation while facilitating usage of technology. Table I illustrate the cluster of respected factors and determinants.

B. Technology Readiness Index 2 (TRI 2)

Technology readiness can be defined as the eagerness of people to accept and adopt the changes in technology which indirectly will incorporate the technology in their work and life [38]. Meanwhile, the Technology Readiness Index (TRI) is a model in measuring people's tendency to embrace new modern technologies [39]**.** Prior studies have shown the excellent result of this model in finding people's tendency to embracing new technology, especially in an organisation. There are four main dimensions of the model including;

1) Optimism: Optimism refers to a positive approach by people towards the use of new or changes in technology [40]. This indirectly plays as a positive factor in the TRI model.

2) Discomfort: Opposite to optimism, discomfort refers to the negative response of people to any changes upon technology. To emphasise, most people find it is uncomfortable to handle new technology or any changes been made upon the technology, as they find the changes are complicated to keep up. Thus, this plays as a negative factor in the TRI model.

3) Insecurity: Insecurity refers to sceptical behaviour, where people lose trust or do not have any trust in technology [41]. To emphasise, most companies reluctant to implement new technologies as they felt insecure regarding the cost of implementing, plus the future direction of the technology remains uncertain [42].

4) Innovativeness: Innovativeness illustrates the level of innovations that are being embraced by people and organisation towards upon the development of cutting-edge technology [40]. This construct also represents the positive construct in the TRI model.

According to [43], optimism and innovativeness are the two positives construct, while discomfort and insecurity are clustered as the negative construct in the TRI model. Consistent with [43], these positive and negative factors enable researchers to investigate the necessity of implementing new technology upon people's behaviour towards the usage of technology. Positive factors will posit the result of people's attraction to new technology. While negative factor's result will postulate that there might be hinder or delay in the overall acceptance to the company or individual. Most of the organisation instigated the TRI model upon the implementation of new technology in their organisation. This is due to the criteria of the TRI model that established based on the psychological assessment of individual or organisation either they will accept or reject the technology that will be used.

The original establishment of TRI model consists of 36-items and divided into four dimensions for each factor, which is (i) optimism [10-items], (ii) innovativeness [7-items], (iii) discomfort [10-items], and (iv) insecurity [9-items] [44]. However, studies show that TRI has a setback result due to the pace of technology advancement [38], [41]. Therefore, TRI 2 is established due to the prior challenge on the first TRI model [43]. Based on the literature, there is evidence that TRI 2 is more robust and have concise result compared to TRI. Compared to 36-items of TRI, TRI 2 only have 6-items with 4-items on each factor. Therefore, TRI 2 is applicable to be implemented in a survey that measures multiple variables aside from the TRI model. Hence, this study adopted the TRI 2.0 model due to the conciseness and robustness, which can be used across time and technology [43], thus making it suitable to be implemented in blockchain acceptance study. Previously, TRI has been suggested to be integrated with TAM, as the TRI factors can be acted as the positive and negative factor that affects PU and PEoU of TAM [11], [45] [46].

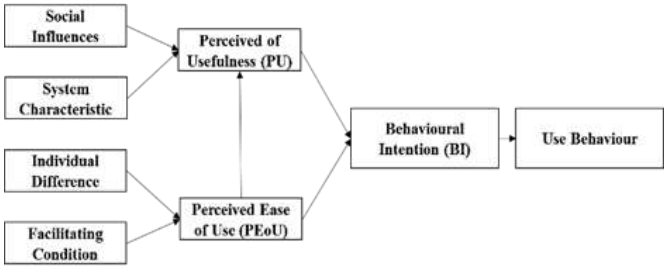


Fig. 1. TAM Theoretically Main Criteria.

TABLE I. THE CLUSTER OF RESPECTED FACTORS AND DETERMINANTS

Variables	**Factors**	**Determinants**
Social Influences	Subjective Norm	PU
System Characteristic	Image	
Individual Differences	Job Relevance, Output Quality, Result Demonstrability	PEoU
Facilitating Conditions	Computer Self-Efficacy, Computer Anxiety, Computer Playfulness	

Additionally, the studies by [46]–[48] shows that TRI construct does significantly related to behavioural acceptance of the individual. Consequently, this study opted to integrate constructs from TAM 3 and TRI 2 in the proposed model of Blockchain-Based Data-sharing Acceptance Model as elaborated in the next section of this paper.

C. Previous Acceptance Study on Blockchain Technology and Cryptocurrency

Technology Acceptance Model had been utilized to study the acceptance and adoption of many kinds of technology, including in the realm of cryptocurrency for individual and target group. Thus, this research considers this model as a suitable model to determine the acceptance and adoption of blockchain. In [49], the authors adopted the Unified Theory of Acceptance and Use of Technology (UTAUT) to investigate elements that are possibly influencing Malaysian banking institutions behavioural intention to adopt blockchain technology. Meanwhile, in [50], the authors proposed a research model which integrates the particular dimension of cryptocurrency into UTAUT and UTAUT2. The integration enables the group to study the factors that influence the acceptance of cryptocurrency in Malaysian individuals' context. The study involved a pilot study of 36 respondent and analysis conducted using PLS-SEM analyses.

In [51], to measure target group intention to use research data sharing system that applied the blockchain-based technologies, researchers had developed a prototype by applying the extended TAM-based model. The authors found that this study gave a basic understanding of the acceptance level on the blockchain-based data sharing; however, no empirical data available to support the finding.

Furthermore, researchers also used TAM to measure blockchain and cryptocurrency acceptance and adoption in [51]–[55]. However, most studies only include few constructs from whether TAM or TRI to study blockchain acceptance; meanwhile, other significant constructs are neglected. The use of incomplete construct might affect the balance of the constructs and scale of TRI compared to the original version of TRI [43], [44]. Hence, this study proposes to integrate TAM3 and TRI2.0, as suggested in the previous study [46], [56].

V. Proposed Model on Evaluation of Blockchain-based Data Sharing Acceptance among Intelligence Community

A. Conceptual Model

Integration of two paradigms between TAM 3 and TRI 2 is considered an established integration model that could deliver the excellent result in previous research. Selected constructs from TAM 3 are Job Relevance for System Characteristics, Computer Self-Efficacy and Computer Anxiety for Individual Difference Factor, Perception of External Controls for Facilitating Condition Factor as shown in Table II.

This selection is made upon the conformity of the target audience, which is the intelligence community.

Bala [35] illustrated that Job Relevance needs to be selected as it is crucial to investigate either people can trust the usage of technology and their belief if the technology improves their life and work. As an example, intelligence community personnel believe that blockchain technology can improve the workflow of information sharing in the intelligence community. The author also highlights the significance of Computer Self-efficacy, Computer Anxiety and Perception of External Control. Computer Self-efficacy enables investigation upon the effect of competency of intelligence personnel upon the acceptance and readiness of blockchain technology implementation.

In comparison, Computer Anxiety examines how intelligence personnel feels upon on the blockchain usage, which will indirectly affect their acceptance of blockchain technology, the PEoU. Perception of External Control will study regarding the available resource and support that can be provided to the intelligence community upon the implementation of blockchain implementation. On the other hand, this study selected all construct from TRI 2, as suggested by [11]. The conceptual model were adapted from previous study by [46], [56] which successfully validated the acceptance of data mining among public university student in Malaysia as it is suitable to be implemented in this study.

Therefore, an overview of the proposed model with respective constructs from TAM 3 and TRI 2 is presented as in Fig. 2.

The established model can be used in investigating an individual's acceptance and readiness for the implementation of blockchain in the intelligence community. Based on the constructed model, a set of hypotheses is developed.

TABLE II. SELECTED CONSTRUCT

Variables	**Factors**	**Determinants**
System Characteristic	Job Relevance	PU
Individual Difference	Computer Self-Efficacy, Computer Anxiety	PEoU
Facilitating Conditions	Perception of External Control	

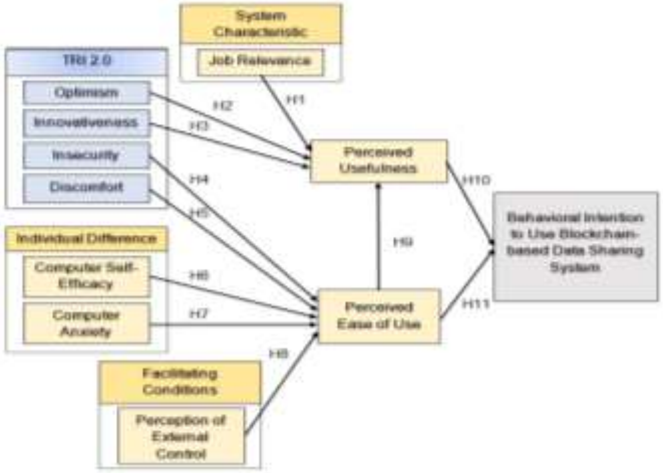


Fig. 2. Overview of the Proposed Conceptual Model with Respective Constructs from TAM 3 and TRI 2.

B. Hypothesis Development

This study suggests the following hypothesis in exploring the influence of the variable to blockchain-based data sharing acceptance and readiness in the intelligence community. Hence, this study hypothesises that:

H1: Job relevance has a positive influence on the perceived usefulness of the blockchain-based data-sharing system.

H2: Optimism has a positive influence on the perceived usefulness of the blockchain-based data-sharing system.

H3: Innovativeness has a positive influence on the perceived usefulness of the blockchain-based data-sharing system.

H4: Insecurity has a negative influence on the perceived ease of use of the blockchain-based data-sharing system.

H5: Discomfort has a negative influence on the perceived ease of use of the blockchain-based data-sharing system.

H6: Computer self-efficacy has a positive influence on the perceived ease of use of the blockchain-based data-sharing system.

H7: Computer anxiety has a negative influence on the perceived ease of use of the blockchain-based data-sharing system.

H8: Perception of external control has a positive influence on the perceived ease of use of the blockchain-based data-sharing system.

H9: Perceived ease of use has a positive influence on the perceived usefulness of the blockchain-based data-sharing system.

H10: Perceived usefulness has a positive influence on the behavioural intention to use the blockchain-based data-sharing system.

H11: Perceived ease of use has a positive influence on the behavioural intention to use the blockchain-based data-sharing system.

VI. METHODOLOGY

A. Instrument Development

This study adopts a quantitative deductive approach of primary data collection using the survey questionnaire. The previous study by [35] [43] [46] and [56] was adapted and tailored accordingly to suit this study. The survey instrument that was developed consisted of 56 questions and divided by two sections include first section; Demographic Information, Authentication and Blockchain Knowledge, and second section; Technology Acceptance Model and Technology Readiness Index. Demographic information comprises the relevant information of the respondents including age, gender, level of education, work experience in the intelligence community, knowledge in authentication system and knowledge on the blockchain application. The questionnaire is measure by 7-points Likert scale in which (1) Strongly Disagree, (2) Quite Disagree, (3) Slightly Disagree, (4) Neutral, (5) Slightly Agree, (6) Quite Agree and (7) Strongly Agree. The questionnaire was validated through a pre-test conducted with 3 respondents from the intelligence community, 2 experts in the blockchain industry and 2 experts in academics to validate the accuracy of the items.

B. Selection of Respondent

This pilot study applies purposive sampling among intelligence personnel from the intelligence community in Malaysia. For the sampling, we used purposive sampling that also referred to as judgement sampling. Participants were selected based on the qualities that the participant holds according to the pre-determined specific criteria [57]. Purposive sampling is commonly used in study using TAM as found in [49] [58] to meet specific criteria of the respondent that is vital in meeting the objectives of the study. Furthermore, the purposive sampling used in a under researched area such as in intelligence community mainly because of the closeness and confidentiality of intelligence practices which made the target population normally reluctant to participate and the sample was chosen exhibit must possess knowledge and experience in intelligence and information sharing, as well as awareness on latest intelligence structures and communication networks [59]. In this study, the sample must work in an intelligence organisation and experienced in the data-sharing system. Targeted respondent is selected and interviewed using a questionnaire where knowledge on authentication and blockchain application is surveyed in the earlier section of the questionnaire. About 35 survey questionnaires were distributed during the interview, and 30 data which meet the criteria were used in the analysis.

VII. RESULT AND FINDINGS

A. Demographic Profile

In total, 30 respondents of this study consist of 23 (76.67%) males and 7 (23.33%) are females. Age distribution of respondent with the majority of 21-30 years old with the sum of 18 (60.00%) respondents, followed by 31-40 years old with 8 respondents (26.67%), 41-50 years old 3 (10.00%) respondents and 51-60 years old 1 (3.33%) respondent. Majority of respondent qualified with bachelor’s degree level of education 19 (63.33%) followed by master’s degree and secondary school qualification of Sijil Pelajaran Malaysia with both 4 (13.33%) respondents, meanwhile for diploma 2 (6.67%) and Doctor of Philosophy (PhD) with 1 (3.33%) respondent. 9 (30%) respondents had 3-5 years working experience in the intelligence community, 7 (23.33%) respondents had less than 3 years working experience in the intelligence community, 7 (23.33%) respondents had 6 to 10 years working experience, meanwhile 4 (13.33%) respondents with more than 16 years’ experience and 3 (10%) respondents with 11-15 years’ experience respectively. To gauge respondent’s knowledge on the authentication system, result from related question shows 23 (76.67%) had knowledge in authentication meanwhile 7 (23.33%) possess no knowledge on authentication system. Data distribution of knowledge on blockchain application shows that 16 (53.33%) had knowledge of blockchain applications; meanwhile, 14 (46.67%) respondents had no knowledge of blockchain application. The indication of high percentage in knowledge about blockchain and authentication system provides credibility of the respondent in answering the questionnaire of this study. The overall demographic information of the respondent is shown in Table III.

B. Reliability and Validity Test

To analyse the reliability and normality, partial least squares structural equation modelling (PLS-SEM) analysis was done by applying Smart PLS 3 software. Based on the model and reference from previous literature, this model is designed and evaluated using a reflective measurement model. The measurement model is assessed by evaluating Internal consistency that includes Cronbach's alpha and composite reliability, Convergent validity that includes indicator reliability and average variance extracted. Also, this research includes the discriminant validity as proposed by [60]. In ensuring the consistency of a measuring instrument, reliability and normality testing is required. Satisfactory level of validity and reliability is required before a significant relationship in the structural model is evaluated [61].

TABLE III. DEMOGRAPHIC INFORMATION (N = 30)

Demographic Criteria	**Frequency**	**Percentage**
Gender		
Male	23	76.67
Female	7	23.33
Age		
21-30	18	60.00
31-40	8	26.67
41-50	3	10.00
51-60	1	3.33
Level of education		
SPM	4	13.33
Diploma	2	6.67
Bachelor's degree	19	63.33
Master's Degree	4	13.33
PhD	1	3.33
Work experience in intelligence community:		
< 3 years	7	23.33
3 – 5 years	9	30.00
6 – 10 years	7	23.33
11 – 15 years	3	10.00
16 years >	4	13.33
Knowledge on Authentication System		
Yes	23	76.67
No	7	23.33
Knowledge on Blockchain Applications		
Yes	16	53.33
No	14	46.67

The assessment of the measurement model that was proposed in this research, Cronbach's Alpha for all construct are analysed. Previous studies recommended that the value for Cronbach's Alpha that greater than 0.7 [61] [60] determined as reliable. Table IV shows all Cronbach's Alpha values is above the acceptable level of 0.7, where the lowest value is Perceived Usefulness (0.722), and the highest value is Job Relevance (0.946). For indicator reliability in exploratory research, values between 0.60 and 0.70 are acceptable; meanwhile, reliability value between 0.70 and 0.95 considered satisfactory to good reliability levels [60]. Hence, 20 indicators with values below 0.6 are eliminated from the original 60 indicators in this study.

After insignificant indicators were eliminated for the model, the composite reliability is evaluated to determine internal consistency. for the composite reliability, the expected minimum level is above 0.70 [60]. As per Table IV, the value of the composite reliability ranged from 0.82 to 0.971. These values are above the recommended acceptable value above 0.70, demonstrating reliability. To assess the convergent validity, the AVE value is evaluated. Convergent validity refers to the theory that indicators for a specific construct are at least moderately correlated between the indicators of constructs [61]. As per Table IV, the AVE value recorded is above 0.5, demonstrate that all the AVE value is satisfactory and reflect that the constructs explain more than half of the indicator's variance [61]. Next, the assessment is done on the discriminant validity. Discriminant validity refers to the extent to which a particular construct varies from others [61]. In this study, as per Table V the discriminant validity is assessed by using the Heterotrait-Monotrait Ratio (HTMT). The author in [61] suggested that HTMT value below 0.8 indicates conceptually different construct. Table V indicates that all the HTMT value is below 0.8 indicates discriminant validity in this study.

TABLE IV. RELIABILITY AND NORMALITY TEST

Variable	**Cronbach's Alpha**	**Composite Reliability**	**AVE**
Behavioural Intention to Use (BIU)	0.836	0.871	0.631
Computer Anxiety (CANX)	0.941	0.971	0.944
Computer Self-Efficacy (CSE)	0.856	0.901	0.696
Discomfort (DISC)	0.807	0.851	0.538
Innovativeness (INN)	0.861	0.877	0.549
Insecurity (INS)	0.787	0.844	0.576
Optimism (OPT)	0.812	0.869	0.575
Perception of External Control (PEC)	0.756	0.820	0.540
Perceived Ease of Use (PEoU)	0.834	0.884	0.658
Perceived Usefulness (PU)	0.722	0.828	0.546
Job relevance (REL)	0.946	0.961	0.860

TABLE V. HETEROTRAIT-MONOTRAIT RATIO (HTMT)

	BIU	CANX	CSE	DISC	INN	INS	OPT	PEC	PEoU	PU	REL
BIU											
CANX	0.094										
CSE	0.299	0.275									
DISC	0.277	0.430	0.646								
INN	0.364	0.467	0.668	0.475							
INS	0.384	0.450	0.318	0.796	0.587						
OPT	0.164	0.195	0.212	0.312	0.378	0.249					
PEC	0.345	0.231	0.757	0.602	0.632	0.406	0.354				
PEoU	0.237	0.138	0.297	0.291	0.307	0.371	0.215	0.401			
PU	0.300	0.270	0.253	0.353	0.288	0.269	0.788	0.288	0.255		
REL	0.139	0.346	0.137	0.264	0.276	0.291	0.244	0.335	0.104	0.714	

C. Structural Model Analysis

Bootstrapping procedure is used in this study to evaluate the significance level of the partial least square estimation [62]. As recommended in reference [60], this study use bootstrapping procedure using 5000 subsamples. Fig. 3 and Table VI shows the values of the path coefficients and R-squared of the structural model.

As per our finding that shown in the structural model result in Fig. 3 and Table VI, consistent with H1, job relevance has a positive influence on the perceived usefulness of blockchain-based data-sharing system with a path coefficient of 0.453. As hypothesised in H2, optimism has a positive influence on the perceived usefulness of blockchain-based data-sharing system with a path coefficient of 0.561. As in H3, the hypothesis is not significant as innovativeness has a negative influence on the perceived usefulness of blockchain-based data-sharing system with a path coefficient of -0.238. For H4, the hypothesis is significant as insecurity has a negative influence on the perceived ease of use of the blockchain-based data-sharing system with a path coefficient value of -0.428.

As in H5, the hypothesis is not significant as the discomfort has a positive influence on the perceived ease of use of the blockchain-based data-sharing system with a path coefficient of 0.018. As hypothesised in H9, perceived ease of use has a positive influence on the perceived usefulness of blockchain-based data-sharing system with a path coefficient of 0.051. The rest of hypothesis is not significant as the result shows a contrast value compared to an early hypothesis in H6, computer self-efficacy has a negative influence on the perceived ease of use of the blockchain-based data-sharing system with a path coefficient of -0.162, in H7, computer anxiety has a positive influence on the perceived ease of use of the blockchain-based data-sharing system with a path coefficient of 0.295. In H8, perception of external control has a negative influence on the perceived ease of use of the blockchain-based data-sharing system with a path coefficient of -0.165, in H10, perceived usefulness has a negative influence on the behavioural intention to use blockchain-based data sharing system with a path coefficient of -0.193 and in H11, perceived ease of use has a negative influence on the behavioural intention to use blockchain-based data sharing system with a path coefficient of -0.011.In order to support the hypothesized paths, as per reference [60], the t values need to be significant at 1.65 (significance level = 0.05), or 2.33 (significance level = 0.01). Based on the result, H1, H2, and H4 are supported meanwhile H3, H5, H6, H7, H8, H9, H10 and H11 are not supported.

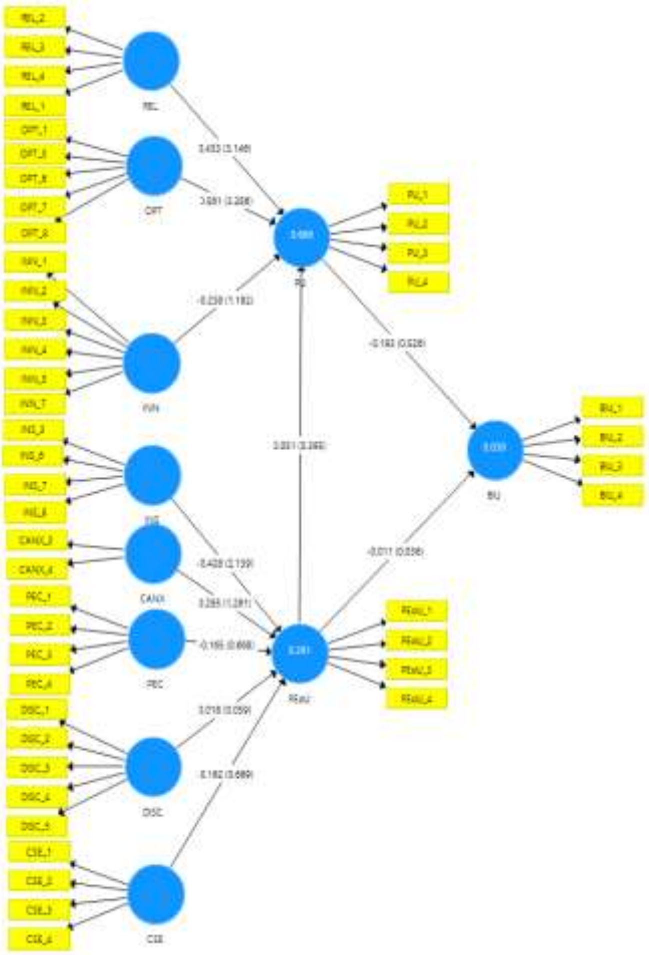


Fig. 3. Structural Model Results.

TABLE VI. RESULT OF STRUCTURAL MODEL TESTING

	Path	Path coefficient	*p*-value	*t*-value	Findings
H1	REL → PU	0.453	0.000*	3.355	Supported
H2	OPT → PU	0.561	0.000*	3.392	Supported
H3	INN → PU	-0.238	0.096	1.305	Not supported
H4	INS → PEoU	-0.428	0.028*	1.917	Supported
H5	DISC → PEoU	0.018	0.477	0.057	Not supported
H6	CSE → PEoU	-0.162	0.257	0.654	Not supported
H7	CANX → PEoU	0.295	0.110	1.225	Not supported
H8	PEC → PEoU	-0.165	0.263	0.636	Not supported
H9	PEoU → PU	0.051	0.351	0.382	Not supported
H10	PU → BIU	-0.193	0.305	0.509	Not supported
H11	PEoU → BIU	-0.011	0.486	0.036	Not supported

*Significant at the 0.05 Level.

Next, the determination of coefficient or R^2 value is carried out. The R^2 value shows the amount of variance in endogenous constructs which the exogenous constructs may describe [63]. The R^2 ranges are from 0 to 1, with higher levels show higher predictive accuracy [60]. In [64], the authors stated R^2 values for endogenous latent variables in the structural model of 0.75 is substantial, 0.50 as moderate and or 0.25 as weak. Fig. 3 shows that the accuracy of the endogenous constructs PU predicted at (68.6 %) and PEOU at (29.1 %), which meant their associated independent variables could explain both dependent variables. The R^2 for BIU is at (3.8 %) indicating weak predictive accuracy.

VIII. DISCUSSION

This study proposed a conceptual model of acceptance and readiness on the blockchain-based data-sharing system for the intelligence community based on constructs in Technology Acceptance Model 3 (TAM 3) and Technology Readiness Level 2 (TRI 2). We achieved our objective of this study to integrate constructs from TAM 3 and TRI 2 in the proposed model of Blockchain-Based Data-sharing Acceptance Model to explore behavioural intention to use the blockchain-based data-sharing system.

This study validates the reliability and validity of the proposed acceptance model using a pilot study conducted among the respondents. Prior to that, a pre-test was conducted to validate the questionnaire survey used in this study with validation from representatives from the respondent group and subject matter expert feedback. From the initial 60 indicators, 20 indicators were removed from the pilot questionnaire due to unsatisfactory level of below 0.6, resulting in the remaining 40 indicators used in further analysis. The reliability and validity of the model are satisfactory and suitable based on the Cronbach's Alpha, Composite Reliability, and Average Variance Extracted (AVE) and discriminant validity based on heterotrait-monotrait ratio (HTMT).

However, the analysis of structural model testing is limited due to the small sample size used in this pilot study. This limitation is similar to other study using a small sample size in investigating acceptance and intention to use blockchain and cryptocurrency as in [50], [51]. According to [60], by using the rule of thumb, the minimum sample size must be 10 times the maximum number of arrowheads in the model. In this model, PEoU has the maximum number of arrowheads pointing to the variable with 5 arrowheads. Hence, at least 45 observations are needed to achieve a statistical power of 80% for at least 0.25 R^2 values detected with a 5% probability of error [60]. In this study, three hypotheses are supported including on job relevance has a positive influence on the perceived usefulness of the blockchain-based data-sharing system, optimism has a positive influence on the perceived usefulness of blockchain-based data-sharing system and insecurity has a negative influence on the perceived ease of use of the blockchain-based data-sharing system.

IX. CONCLUSION AND FUTURE WORK

This paper elaborated and discussed regarding blockchain technology acceptance in intelligence community using the case of the proposed blockchain-based data-sharing system in the intelligence community. As known, the intelligence community relies on accurate and precise information for country security purposes. Thus, blockchain technology is proposed to be integrated into the intelligence community data sharing system due to its capability in managing access control and authentication automatically. Blockchain is also proven to be a brilliant solution in ensuring data integrity that is vital for the intelligence community-related data. However, since that blockchain technology is still new, the readiness and acceptance level of the intelligence community upon blockchain technology implementation is yet to be discovered. Thus, this paper survey about blockchain technology and proposes a pilot study by integrating a reliable model in investigating the intelligence community readiness and acceptance upon blockchain technology usage. The model is established based on constructs from TAM 3 and TRI 2. The establishment of the integrated model derived by the effectiveness that was proven by other researchers in their previous work that we obtained from literature reviews. This study is an ongoing work of implementing TAM 3 and TRI 2 for blockchain technology readiness and acceptance in the intelligence community of Malaysia.

This study concludes that the acceptance model can be used in investigating behavioural intention to use the blockchain-based data-sharing system in the intelligence community. The awareness and knowledge in blockchain technology among the respondent shall be enriched via training and education to increase the level of acceptance and readiness of such technology. Future work may include full-scale survey based on the recommended sample size and involvement of different agencies from the intelligence community context. This will provide reliable data that could serve as a source of reference for the development of government policy for blockchain implementation, especially in the intelligence community environment.

ACKNOWLEDGMENTS

This research is fully supported by National Defence University of Malaysia (UPNM) under Short Grant UPNM/2020/GPJP/ICT/6 and Center of Research and Innovation Management. The authors fully acknowledged the Malaysia Ministry of Higher Education (MOHE) and UPNM for the approved fund, which makes this important research viable and effective.

REFERENCES

[1] W. N. Wan Muhamad et al., "Enhance multi-factor authentication model for intelligence community access to critical surveillance data," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019, vol. 11870 LNCS, pp. 560–569, doi: 10.1007/978-3-030-34032-2_49.

[2] J. Schmid, "Technology and the Intelligence Community," in Advanced Sciences and Technologies for Security Applications, 2018, pp. 39–53.

[3] W. J. Lahneman, "Knowledge-sharing in the intelligence community after 9/11," Int. J. Intell. CounterIntelligence, vol. 17, no. 4, pp. 614–633, 2004, doi: 10.1080/08850600490496425.

[4] S. N. Q. S. Mohamed and M. Yaacob, "Understanding the Intelligence Failure and Information Sharing in Handling Terrorism among Intelligence Community," Int. J. Acad. Res. Bus. Soc. Sci., vol. 9, no. 9, pp. 1201–1213, 2019, doi: 10.6007/ijarbss/v9-i9/6414.

[5] J. W. Crampton, "Collect it all: national security, Big Data and governance," GeoJournal, vol. 80, no. 4, pp. 519–531, 2015, doi: 10.1007/s10708-014-9598-y.

[6] S. S. De Matas and B. P. Keegan, "An exploration of research information security data affecting organizational compliance," Data Br., vol. 21, pp. 1864–1871, 2018, doi: 10.1016/j.dib.2018.11.002.

[7] C. Lin, D. He, X. Huang, K. R. Choo, and A. V Vasilakos, "BSeIn : A blockchain-based secure mutual authentication with fine-grained access control system for industry 4 . 0 ☆," J. Netw. Comput. Appl., vol. 116, no. March, pp. 42–52, 2018, doi: 10.1016/j.jnca.2018.05.005.

[8] P. N. Mahalle, B. Anggorojati, N. R. Prasad, and R. Prasad, "Identity Authentication and Capability Based Access Control (IACAC) for the Internet of Things," J. Cyber Secur. Mobil., vol. 1, pp. 309–348, 2013.

[9] A. Ouaddah, A. A. Elkalam, and A. A. Ouahman, "FairAccess : a new Blockchain-based access control framework for the Internet of Things," no. February, pp. 5943–5964, 2017, doi: 10.1002/sec.1748.

[10] X. Xu et al., "A Taxonomy of Blockchain-Based Systems for Architecture Design," 2017 IEEE Int. Conf. Softw. Archit., pp. 243–252, 2017, doi: 10.1109/ICSA.2017.33.

[11] N. Larasati, "Technology Readiness and Technology Acceptance Model in New Technology Implementation Process in Low Technology SMEs," Int. J. Innov. Manag. Technol., vol. 8, no. 2, pp. 113–117, 2017, doi: 10.18178/ijimt.2017.8.2.713.

[12] O. Novo, "Blockchain Meets IoT: An Architecture for Scalable Access Management in IoT," IEEE Internet Things J., vol. 5, no. 2, pp. 1184–1195, 2018, doi: 10.1109/JIOT.2018.2812239.

[13] T. Tuan, A. Dinh, R. Liu, M. Zhang, and G. Chen, "Untangling Blockchain : A Data Processing View of Blockchain Systems," IEEE Trans. Knowl. Data Eng., vol. 30, no. 7, pp. 1366–1385, 2018, doi: 10.1109/TKDE.2017.2781227.

[14] C. Tankard, "Big data security," Netw. Secur., vol. 2012, no. 7, pp. 5–8, 2012, doi: 10.1016/S1353-4858(12)70063-6.

[15] N. Kshetri, "Big data׳s impact on privacy, security and consumer welfare," Telecomm. Policy, vol. 38, no. 11, pp. 1134–1145, Dec. 2014, doi: 10.1016/j.telpol.2014.10.002.

[16] E. Gaetani, L. Aniello, R. Baldoni, F. Lombardi, A. Margheri, and V. Sassone, "Blockchain-based database to ensure data integrity in cloud computing environments," in CEUR Workshop Proceedings, 2017, vol. 1816, pp. 146–155.

[17] O. Alphand et al., "IoTChain : A Blockchain Security Architecture for the Internet of Things," 2018 IEEE Wirel. Commun. Netw. Conf., pp. 1–6, 2018.

[18] W. Zhang et al., "Blockchain-Based Distributed Compliance in Multinational Corporations' Cross-Border Intercompany Transactions," in Future of Information and Communication Conference (FICC), 2019, no. July, pp. 304–320, doi: 10.1007/978-3-030-03405-4_20.

[19] R. Beck, J. S. Czepluch, N. Lollike, and S. Malone, "Blockchain – The Gateway to Trust-free Cyrptographic Transactions," Twenty-Fourth Eur. Conf. Inf. Syst. (ECIS), İstanbul,Turkey, vol. 6, no. May, pp. 4013–4027, 2016.

[20] J. P. Es-Samaali, H., Outchakoucht, A., & Leroy, "A Blockchain-based Access Control for Big Data," J. Comput. Networks Commun. Secur. Internet Things J., vol. 5, no. 7, p. 137, 2017, doi: 10.1109/JIOT.2018.2812239.

[21] M. Milutinovic, W. He, H. Wu, and M. Kanwal, "Proof of Luck: An efficient blockchain consensus protocol," in SysTEX 2016 - 1st Workshop on System Software for Trusted Execution, colocated with ACM/IFIP/USENIX Middleware 2016, 2016, pp. 2–7, doi: 10.1145/3007788.3007790.

[22] . Karamitsos, M. Papadaki, and N. B. Al Barghuthi, "Design of the Blockchain Smart Contract: A Use Case for Real Estate," J. Inf. Secur., vol. 09, no. 03, pp. 177–190, 2018, doi: 10.4236/jis.2018.93013.

[23] X. Xu, Q. Lu, Y. Liu, L. Zhu, H. Yao, and A. V. Vasilakos, "Designing blockchain-based applications a case study for imported product traceability," Futur. Gener. Comput. Syst., vol. 92, pp. 399–406, 2019, doi: 10.1016/j.future.2018.10.010.

[24] P. Novotny et al., "Permissioned blockchain technologies for academic publishing," Inf. Serv. Use, vol. 38, no. 3, pp. 159–171, 2018, doi: 10.3233/ISU-180020.

[25] X. Cheng and F. Chen, "Design of a Secure Medical Data Sharing Scheme Based on Blockchain," J. Med. Syst., vol. 44, no. 2, pp. 1–11, 2020.

[26] X. Yue, H. Wang, D. Jin, M. Li, and W. Jiang, "Healthcare Data Gateways : Found Healthcare Intelligence on Blockchain with Novel Privacy Risk Control," J. Med. Syst., 2016, doi: 10.1007/s10916-016-0574-6.

[27] C. Service and P. Via, "MeDShare : Trust-less Medical Data Sharing Among," IEEE Access, vol. 5, pp. 1–10, 2017, doi: 10.1109/ACCESS.2017.2730843.

[28] R. Xu, Y. Chen, E. Blasch, and G. Chen, "BlendCAC : A BLockchain-ENabled Decentralized Capability-based Access Control for IoTs," 2018 IEEE Int. Conf. Internet Things IEEE Green Comput. Commun. IEEE Cyber, Phys. Soc. Comput. IEEE Smart Data, pp. 1027–1034, 2018, doi: 10.1109/Cybermatics.

[29] K. Naerland, C. Müller-Bloch, R. Beck, and S. Palmund, "Bill of Lading on Blockchain Blockchain to Rule the Waves - Nascent Design Principles for Reducing Risk and Uncertainty in Decentralized Environments," Proc. Int. Conf. Inf. Syst., pp. 1–16, 2017.

[30] A. Tarhini, N. A. G. Arachchilage, R. Masa'deh, and M. S. Abbasi, "A Critical Review of Theories and Models of Technology Adoption and Acceptance in Information System Research," Int. J. Technol. Diffus., vol. 6, no. 4, pp. 58–77, 2015, doi: 10.4018/ijtd.2015100104.

[31] C. E. Porter and N. Donthu, "Using the technology acceptance model to explain how attitudes determine Internet usage: The role of perceived access barriers and demographics," J. Bus. Res., vol. 59, no. 9, pp. 999–1007, 2006, doi: 10.1016/j.jbusres.2006.06.003.

[32] A. Indrati et al., "Validity of the technology acceptance model (tam): a sensemaking perspective," Ijms, vol. 6, no. 1, pp. 99–120, 2012, doi: 10.5897/AJBM10.1398.

[33] Q. L. Chen and Z. H. Zhou, "Unusual formations of superoxo heptaoxomolybdates from peroxo molybdates," Inorg. Chem. Commun., vol. 67, no. 3, pp. 95–98, 2016, doi: 10.1016/j.inoche.2016.03.015.

[34] V. Venkatesh and F. D. Davis, "Theoretical extension of the Technology Acceptance Model: Four longitudinal field studies," Manage. Sci., vol. 46, no. 2, pp. 186–204, 2000, doi: 10.1287/mnsc.46.2.186.11926.

[35] V. Venkatesh and H. Bala, "Technology Acceptance Model 3 and a Research Agenda on Interventions," Decis. Sci., vol. 39, no. 2, pp. 273–315, 2008.

[36] D. J. McFarland and D. Hamilton, “Adding contextual specificity to the technology acceptance model,” Comput. Human Behav., vol. 22, no. 3, pp. 427–447, May 2006, doi: 10.1016/j.chb.2004.09.009.

[37] P. Lai, “The Literature Review Of Technology Adoption Models And Theories For The Novelty Technology,” J. Inf. Syst. Technol. Manag., vol. 14, no. 1, pp. 21–38, Apr. 2017, doi: 10.4301/S1807-17752017000100002.

[38] K. M. Kuo, C. F. Liu, and C. C. Ma, “An investigation of the effect of nurses’ technology readiness on the acceptance of mobile electronic medical record systems,” BMC Med. Inform. Decis. Mak., vol. 13, no. 1, pp. 1–14, 2013, doi: 10.1186/1472-6947-13-88.

[39] A. Caison, D. Bulman, S. Pai, and D. Neville, “Exploring the technology readiness of nursing and medical students at a Canadian University,” J. Interprof. Care, vol. 22, no. 3, pp. 283–294, 2008, doi: 10.1080/13561820802061809.

[40] A. Parasuraman and Charles L. Colby, Techno-Ready Marketing: How and Why Your Customers Adopt Technology. NY, USA: The Free Press New York, 2007.

[41] J. S. C. Lin and P. L. Hsieh, “Refinement of the technology readiness index scale: A replication and cross-validation in the self-service technology context,” J. Serv. Manag., vol. 23, no. 1, pp. 34–53, 2012, doi: 10.1108/09564231211208961.

[42] S. Shin and W. J. Lee, “The effects of technology readiness and technology acceptance on NFC mobile payment services in Korea,” J. Appl. Bus. Res., vol. 30, no. 6, pp. 1615–1626, 2014.

[43] A. Parasuraman and C. L. Colby, “An Updated and Streamlined Technology Readiness Index: TRI 2.0,” J. Serv. Res., vol. 18, no. 1, pp. 59–74, 2015, doi: 10.1177/1094670514539730.

[44] A. Parasuraman, “Technology Readiness Index (TRI): A Multipleitem Scale To Measure Readiness To Embrace New Technologies,” J. Serv. Res., vol. 2:307, no. May, 2000.

[45] C.-H. Lin, H.-Y. Shih, and P. J. Sher, “Integrating technology readiness into technology acceptance: The TRAM model,” Psychol. Mark., vol. 24, no. 7, pp. 641–657, Jul. 2007, doi: 10.1002/mar.20177.

[46] M. Wook, Z. M. Yusof, and M. Zakree Ahmad Nazri, “The Acceptance of Educational Data Mining Technology among Students in Public Institutions of Higher Learning in Malaysia,” Int. J. Futur. Comput. Commun., vol. 4, no. 2, pp. 112–117, 2015, doi: 10.7763/ijfcc.2015.v4.367.

[47] J. S. C. Lin and H. C. Chang, “The role of technology readiness in self-service technology acceptance,” Manag. Serv. Qual., vol. 21, no. 4, pp. 424–444, 2011, doi: 10.1108/09604521111146289.

[48] Y. Yi, L. L. Tung, and Z. Wu, “Incorporating Technology Readiness (TR) Into TAM : Are Individual Traits Important to Understand Technology Acceptance ?,” Digit 2003 Proc., pp. 1–27, 2003.

[49] H. Yusof et al., “Behavioral Intention to Adopt Blockchain Technology: Viewpoint of the Banking Institutions in Malaysia,” Int. J. Adv. Sci. Res. Manag., vol. 3, no. 10, pp. 1–6, 2018, [Online]. Available: www.ijasrm.com.

[50] Y. C. Yeong, K. S. Kalid, and S. K. Sugathan, “Cryptocurrency acceptance: A case of Malaysia,” Int. J. Eng. Adv. Technol., vol. 8, no. 5, pp. 28–38, 2019, doi: 10.35940/ijeat.E1004.0585C19.

[51] A. K. Shrestha and J. Vassileva, “User acceptance of usable blockchain-based research data sharing system: An extended TAM-based study,” Proc. - 1st IEEE Int. Conf. Trust. Priv. Secur. Intell. Syst. Appl. TPS-ISA 2019, pp. 203–208, 2019, doi: 10.1109/TPS-ISA48467.2019.00033.

[52] L. Wanitcharakkhakul and S. Rotchanakitumnuai, “Blockchain technology acceptance in electronic medical record system,” Proc. Int. Conf. Electron. Bus., vol. 2017-Decem, pp. 53–58, 2017.

[53] C. C. Lee, J. C. Kriscenski, and H. S. Lim, “An Empirical Study Of Behavioral Intention To Use Blockchain Technology.: Sistema de descoberta para FCCN,” J. Int. Bus. Discip., vol. 14, no. 1, pp. 1–21, 2019, [Online]. Available: https://eds.b.ebscohost.com/eds/pdfviewer/pdfviewer?vid=4&sid=c4c41200-8ffa-4706-9562-33088ffd69aa%40pdc-v-sessmgr02.

[54] S. Supranee, S. Rotchanakitumnuai, and R. Siriluck, “The Acceptance of the Application of Blockchain Technology in the Supply Chain Process of the Thai Automotive Industry,” 2017, [Online]. Available: http://aisel.aisnet.org/iceb2017http://aisel.aisnet.org/iceb2017/30.

[55] I. Roussou, E. Stiakakis, and A. Sifaleras, “An empirical study on the commercial adoption of digital currencies,” Inf. Syst. E-bus. Manag., vol. 17, no. 2–4, pp. 223–259, 2019, doi: 10.1007/s10257-019-00426-7.

[56] M. Wook, S. Ismail, N. M. M. Yusop, S. R. Ahmad, and A. Ahmad, “Identifying priority antecedents of educational data mining acceptance using importance-performance matrix analysis,” Educ. Inf. Technol., vol. 24, no. 2, pp. 1741–1752, 2019, doi: 10.1007/s10639-018-09853-4.

[57] I. Etikan, “Comparison of Convenience Sampling and Purposive Sampling,” Am. J. Theor. Appl. Stat., vol. 5, no. 1, p. 1, 2016, doi: 10.11648/j.ajtas.20160501.11.

[58] Y. Malhotra and D. F. Galletta, “Extending the Technology Acceptance Model to account for social influence: Theoretical bases and empirical validation,” Proc. Hawaii Int. Conf. Syst. Sci., vol. 00, no. c, p. 5, 1999, doi: 10.1109/hicss.1999.772658.

[59] J. Carter, “Inter-organizational relationships and law enforcement information sharing,” no. May, 2015, doi: 10.1080/0735648X.2014.927786.

[60] Hair, G. T. Hult, C. Ringle, and M. Sarstedt, A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM) - Joseph F. Hair, Jr., G. Tomas M. Hult, Christian Ringle, Marko Sarstedt. Los Angeles: SAGE Publications, Inc. Printed, 2017.

[61] A. Leguina, “A primer on partial least squares structural equation modeling (PLS-SEM),” Int. J. Res. Method Educ., vol. 38, no. 2, pp. 220–221, 2015, doi: 10.1080/1743727x.2015.1005806.

[62] W. W. Chin, Handbook of Partial Least Squares. 2010.

[63] M. Sarstedt, C. M. Ringle, and J. F. Hair, Handbook of Market Research, no. September. 2017.

[64] J. F. Hair, C. M. Ringle, and M. Sarstedt, “PLS-SEM: Indeed a silver bullet,” J. Mark. Theory Pract., vol. 19, no. 2, pp. 139–152, 2011, doi: 10.2753/MTP1069-6679190202.

Predicting Mental Illness using Social Media Posts and Comments

Mohsin kamal[1], Saif Ur Rehman khan[2], Shahid Hussain[3]
Anam Nasir[4], Khurram Aslam[5], Subhan Tariq[6], Mian Farhan Ullah[7]*
Department of Computer Science, COMSATS University, Islamabad, Pakistan[1, 2, 3, 4, 6]
Department of Computer Science, University of Oregon, USA[5]
Electrical Engineering Department, Wah Engineering College, University of Wah, Wah Cantt, Pakistan[7]

***Abstract*—From the last decade, a significant increase of social media implications could be observed in the context of e-health. The medical experts are using the patient's post and their feedbacks on social media platforms to diagnose their infectious diseases. However, there are only few studies who have leveraged the capabilities of machine learning (ML) algorithms to classify the patient's mental disorders such as Schizophrenia, Autism, and Obsessive-compulsive disorder (OCD) and Post-traumatic stress disorder (PTSD). Moreover, these studies are limited to large number of posts and relevant comments which could be considered as a threat for their effectiveness of their proposed methods. In contrast, this issue is addressed by proposing a novel ML methodology to classify the patient's mental illness on the basis of their posts (along with their relevant comments) shared on the well-known social media platform "Reddit". The proposed methodology is exploit by leveraging the capabilities of widely-used classifier namely "XGBoost" for accurate classification of data into four mental disorder classes (Schizophrenia, Autism, OCD and PTSD). Subsequently, the performance of the proposed methodology is compared with the existing state of the art classifiers such as Naïve Bayes and Support vector machine whose performance have been reported by the research community in the target domain. The experimental result indicates the effectiveness of the proposed methodology to classify the patient data more effectively as compared to the state of the art classifiers. 68% accuracy was achieved, indicating the efficacy of the proposed model.**

Keywords*—*Machine learning; mental disorders; Reddit; Schizophrenia; Autism; OCD; PTSD

I. Introduction

Mental disorders are growing sprightly all over the world and World Health Organization (WHO) predicted that one of four people around the world, at some point in their lives will be afflicted with mental disorders. Furthermore, according to Üstün [1], Depressive disorders will become second largest cause of the global disease burden. Around the world, thousands of Men, women, adults and even children are suffering from mental disorders. Mental illness is a disease that negatively impacts the human behavior and thoughts. Subsequently, disturbing an individual's social and domestic life. A Mental disorder has more than two hundred classified forms, in which some of the more common types are bipolar disorder, schizophrenia, depression, stress, anxiety and dementia. There are symptoms that determine either an individual is suffering from mental illness or not. These symptoms include: strong feelings of anger, dramatic changes in eating and sleeping, excessive fears, feelings of extreme high and low, worries, stress, anxieties and suicidal thoughts [2]. People with serious mental illness have a high mortality rate. The authors identified an excessive mortality rate among an individual of serious mental illness, they listed episodic depression and recurrent depressive disorder. The results showed the mortality rate varies with gender, age, diagnosis and ethnicity [3].

The risk of mental disorder misdiagnosing is decreasing and individuals are being diagnosed on time due to the awareness of symptoms and predominantly on introducing Machine Learning (ML) techniques for the diagnosis of mental disorder. Machine Learning techniques are widely used in the medical field, as they provide high accuracy and effective results. Furthermore, with the use of machine learning techniques, the symptoms of mental disorder are predicted and based on the predictions the individual are informed about the condition of their mental health. Pattern recognition algorithm of ML is well-recognized for the treatment, diagnosis and prediction of complications in the treatment of mental disorder [4]. Machine learning acquired deep data or big data for the best results. ML algorithm such as Decision Tree, Support Vector Machine (SVM), Naïve Bayes (NB), Logistic Regression (LG) and k-Nearest Neighbor (KNN) classifiers are used to analyze the state of mental health of a particular group of individuals. All of the ML techniques performed well by achieving an accuracy up to 85% [5] subject to dataset characteristics, which threat their employment in real world scenarios.

The prediction of mental health from individual's social media posts (e.g. Facebook, Twitter, Instagram and Reddit) is one of the encroachments. Social media is a great source of communication and interaction among people. Where they share their opinions and thoughts with each other, such as (posting their photos, videos and comments) which reflect their feelings, mood and sentiments. Hence, their mood and emotions from their posts and comments can be predicted using machine learning algorithms [6]. In [7] the authors analyze the depression on the data gathered from Facebook posts. According to [8] Instagram is another social network, which is widely used all over the world and daily millions of people share their feelings and opinions on it. The authors predicted the markers of depression from Instagram data. They extracted the statistical features from 43,950 individual Instagram photos, such as, metadata component, color analysis

*Corresponding Author

and algorithmic face detection. The machine learning models well performed and predicted early mental disorder and screening.

In this research work, a methodology is proposed for the identification of the mental illness of a patient via their communication on the social media networks. The data for this study were drawn from a well-known social media platform "Reddit". Though, there are several mental illnesses whose data can be collected from social media network. However, data is mine (i.e. post and comments) of certain widely-known mental illness classes such as Schizophrenia, Autism, OCD and PTSD. Furthermore, the proposed model evaluated in terms of accuracy, precision, recall and f-measure. Two research questions are formulated to conduct this study:

Research Question 1: Is the proposed methodology (utilizing XGBoost) in terms of identifying the metal disorders?

Research Question 2: Which is more accurate ML algorithm for the prediction of mental disorder?

The rest of the paper is organized as follows: Section 2 presents the related work, Section 3 discusses about the proposed methodology. In Section 4, presents results and discussion. Finally, Section 5 presents the conclusion.

II. Related Work

Many techniques have been developed and adopted by different researchers for automated prediction of mental disorders such as, schizophrenia, PTSD, OCD and ASD. Schizophrenia is indicated by thought disorder, hallucinations, cognitive deficits and delusions. In the early stage of schizophrenia authentic diagnosis is very important but remains challenging. In study [9] the classification between the control groups and schizophrenic patients was done on basis of magnetic resonance imaging (MRI) using ML. The authors implemented a multimodal classification technique (*structural MRI, Diffusion Tensor Imaging (DTI) and resting state-functional MRI data*) to differentiate drug-naïve first-episode schizophrenic patients from control groups. The authors used feature selection sparse coding (SC) and multi-kernel SVM for the features combination and classification. Similarly, in study [10] analysis of DNA methylation was implemented by the modern machine learning and Bio-conductor Minfi package. Case-controls studies in methylated patterns were successfully detected and highlighted. The classification of schizophrenia patients and healthy control groups was carried out through extracted features of both source-level and sensor-level from EEG signals taped during an auditory oddball task. The results of the study indicated a higher classification accuracy, as the features of both source-level and sensor-level are used together[11]. Some psychiatric diseases such as bipolar and schizophrenia disorder are very contrasted both in etiology and clinical manifestation [12].

Similarly, for the prediction of post-traumatic stress disorder (PTSD) ML also plays a vital role. As, in study [13] machine learning approach was used to forecast the long-term PTSD and risk indicators in soldiers from Afghanistan. The ML approach performed very well and resulted an outstanding performance to analyze the PTSD in soldiers. In [14], PTSD was predicted using supervised ML in order to crack the interchangeable and increasing risk indicator combinations. SVM and 10-fold cross validation were used for PTSD. Furthermore, [15] presents a ML technique to classify between the child PTSD and healthy group. Results were accurate, showing that ML is much reliable and helpful in predicting mental disorders.

Compulsivity disorders such as obsessive-compulsive disorder (OCD) and stimulant use disorder (SUD) are indicated by a loss in behavioral flexibility. As, in [16] OCD and SUD are identified through probabilistic reversal learning (PRL) paradigms. In this work the author applied hierarchical Bayesian model to PRL data of an individual with an OCD, SUD and the healthy controls. According to [17] the effective treatment for OCD is cognitive behavioral therapy (CBT). In this study, the authors gathered resting state functional MRI of patients with OCD before and after 4 weeks of daily intensive CBT. Cross validation was applied to observe the Functional Connectivity (FC) patterns to predict OCD symptoms of an individual. Similarly, in [18] the patients with pediatric OCD ,who acquired to Internet-delivered cognitive behavioral therapy (ICBT) were tested by four machine learning techniques(Random Forest, SVM, linear model and the L1 elastic Net (Lasso)). In study [19], Deep brain stimulation (DBS) treatment was applied on two patients with tractable OCD using automated face analysis. The activation of DBS device and intraoperative implantation was applied on one patient, while the other one was assessed with 3 month post-implantation. Convolutional Neural Network (CNN) was used to quantify positive and negative valence. This technique yield better results in both situations.

Likewise, Autism spectrum disorder (ASD) is a developmental neuropsychiatric disorder characterized by impairments in communication (both verbal and non-verbal), social interaction and restricted, repetitive behavior. According to [20] the diagnosis of ASD is slowed because it requires administration for standardized examination, such as the Autism Diagnostic Observation Schedule (ADOS). Furthermore, it took several hours to analyze 20 to 100 behaviors of an ASD patient through standard approaches of diagnosis. So, to overcome this, the authors applied 17 unique supervised learning methods to classify dataset in five classes. In [21] the authors evaluate video recording from two standard diagnostic instruments in order to design ML classifiers to optimize interpretability, sparsity and accuracy. In [22], the classification of the low functioning of children aged between 2-4 years was possible. The dataset was classified in two classes ASD vs TD by using ML pattern classification technique. This research shows the maximum classification accuracy with 7 features that belong to the goal-oriented of movement of the body parts.

From the above related work it can be concluded that Machine Learning (ML) plays a significant role in automated prediction of mental disorders. Furthermore, the above studies present SVM and NB as the two most accurate algorithms for the classification of mental illness.

III. Methodology

The proposed model comprises of three phases (Fig. 1). In the first phase, the data were collected from a social network “Reddit”. In the second phase, preprocessing was done on the dataset and then feature selection techniques were applied. Finally, in the third phase the model was trained, and was tested.

A. Data Acquisition

Data collection is the first phase in the proposed model. For the data collection purpose social media platform “Reddit” is chosen, and only clinical posts were selected from “subreddits”. Subreddits are the pages that are made by the users of “Reddit” for specified topics. Posts from four clinical subreddits are collected: r/Schizophrenia, r/Autism, r/OCD and r/PTSD. Furthermore, 1000 posts from each subreddit is collected using the Reddit Application Programming Interface (API: reddit.com/dev/api). After collection of posts, the comments are gathered and store in another csv file.

B. Data Preprocessing

Data pre-processing is the second phase of the proposed model. Pre-processing activities such as tokenization, removal of stop words, punctuation removal, word stemming and vectorization are applied to the collected data to make the data useful for test model. The chain of preprocessing activities is presented in Fig. 2.

1) Tokenization: Tokenization is the first activity. It is a process to break down the text data into smaller units called tokens. The token may be the symbols, words, numbers, phrases or other meaningful elements. The list of these tokens are input to other preprocessing activities.

2) Stop word: The second activity is the removal of stop words. Stop words are the words which are commonly used in sentences and phrases and have no significant information, such as “the, is, am, are, a, an”.

3) Remove punctuations: The third activity is the removal of punctuation. It is the process to remove all punctuations from the data such as: “!”$#%&’()*+,-./:;<=>?@[\]^_`{|}~”, ☺, 😐, ☹ etc.

4) Word stemming: The fourth activity is word stemming. It is the process to compress each word into a common root or base. It involves cutting off the prefixes or suffixes of the word and change it into an inflected word.

In preprocessing, non-alpha characters and spaces are removed. After completing the preprocessing steps, the data convert into the form of vectors.

5) Encoding data: In this activity the target data are encoded in numeric form. As, here a multi-class problem, the available data are categorical data which are in form of string. So, the data are labelled in the form of numeric values, which is understandable for test model.

6) Word vectorization: Word vectorization is another activity in pre-processing phase. It is a process to convert all the stream of text data into numerical feature vectors. As the proposed model only understand the numeric data so, word vectorization is a necessary task in this phase. To convert the text stream of data into numeric data, a famous technique Term Frequency- Inverse Document Frequency (TF-IDF) is being used. TF-IDF is widely used for weighting the text stream to retrieve the information [23]. It implicates, adjusting the frequencies of the data. Frequency depends upon the number of occurrences of a word. If a word occurs several times, then it is assigned a high frequency numerical value. Basically, TF-IDF offers weight for each word. A TF-IDF score is given by the formula:

$$tfidf(t,d,D) = \frac{tf(t,d)}{df(t,d)} \quad (1)$$

A Python library “scikit-learn” is used to apply this scaling and extract features from the data. A maximum of 5000 features is extracted from the data set.

C. Feature Selection

Feature selection is the last activity of the data pre-processing phase. It is an important step to select the significant features from the data set.

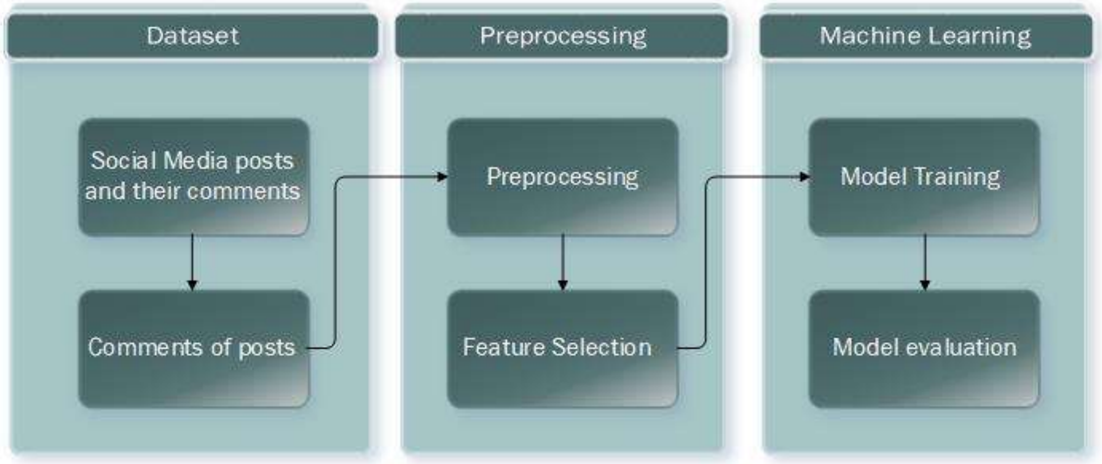


Fig. 1. Graphical Overview of Proposed Model.

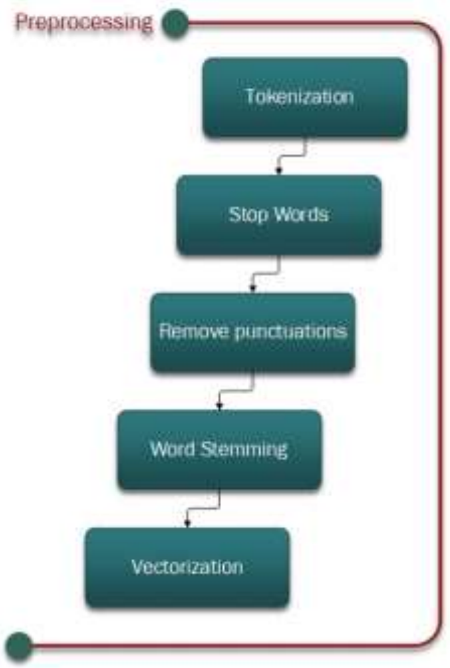


Fig. 2. Steps Involved in Preprocessing Phase.

As test data consist of thousands of comments and those comments consist of millions of word stream, so despite of useful features, this word stream also contain many useless features that increase the processing time and create distortion in the data. Hence, it is necessary to remove these useless features and select the only features that are significantly useful. The Random Forest technique is used for this purpose. It is a subsist of 4-12 hundred decision trees. Each of the decision tree is made over a random extraction of analysis from the dataset and a random selection of features. The decision trees are de-correlated and less prostrate to overfitting hence, improving the purity of node which allow the decrement of impurity from all over the tree and selecting the best feature of the data. An automatic feature selection is used and 1000 best features from the data are selected and passed to machine learning model for the classification.

D. *Machine Learning Model*

In this research work, XGBoost technique is used in order to classify the comments on the posts. XGBoost technique was proposed in 2016, it is a boosting technique to increase the performance and decrease the processing time. In the proposed model it is used to classify the comments on the posts into four classes i.e. r/Schizophrenia, r/Autism, r/OCD and r/PTSD. Furthermore the accuracy of the model is calculated through F-measure, which is a standard statistic calculation for machine learning classifiers. F-measure is the average recall and precision. Recall states the sensitivity of the machine learning model. It defines the ratio of accurately predicted positive observations to the total no. of the observations presented in the actual class. For example, a good schizophrenia classifier should predict maximum comments from the Schizophrenia subreddit.

$$Recall\ = \frac{TP}{(TP + FN)} \tag{2}$$

Similarly, Precision is the total ratio of accurately predicted positive observations to the total number of positive predictive observations. For example, when a classifier labels a comment as from the Schizophrenia subreddit then its prediction must be correct.

$$Precision = \frac{TP}{(TP + TP)} \tag{3}$$

The F-measure is usually preferred over accuracy. It is a harmonic mean of Precision and Recall.

$$F1\ Score\ = 2 * \frac{Recall * Precision}{(Recall + Precision)} \tag{4}$$

The accuracy is calculated as a whole performance of the model.

$$Accuracy = \frac{(TP + TN)}{TP+TN+FP+FN} \tag{5}$$

IV. Results and Discussion

In this research work, the dataset has generated about 32330 comments on 4000 posts on clinical subreddit. In each experiment, K-fold (i.e. K=10) cross validation is performed to analyze the effectiveness of the proposed methodology. Four widely used performance measure, namely Accuracy, Precision, Recall, F-measure is used in each experiment.

A. *Response to RQ-1*

In order to respond RQ-1, first, an experiment is performed to investigate the effectiveness of the proposed methodology employing the XGBoost classifier. Firstly, the training and cross-validation score in terms of accuracy is shown in Fig. 3. The class wise (in terms of OCD, Autism, PTSD, and Schizophrenia) performance of XGBoost in terms of Precision, Recall, and F-measure is shown in Table I, while the average performance in terms of Accuracy is shown in Fig. 3. Besides, the training Log Loss (>0.69) indicate the low risk in prediction and better performance of XGBoost classifier.

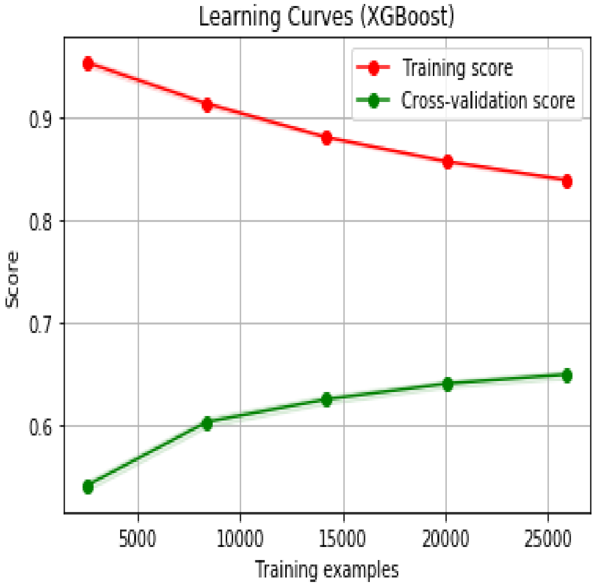


Fig. 3. Learning curve of XGBoost.

Log-loss is also a measurement of the accuracy that includes the concept of probability assurance. As classification accuracy solely is not enough to analyze the strength of the prediction so, log loss is also a significant measure. Basically, it is a Cross entropy concerning the true labels and predicted probabilities. It is calculated as:

$$Log\ Loss - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} y_{i,j} log(p_{ij}) \qquad (6)$$

Where N is the number of samples and M is the number of classes, yij is true label and pij is predicted probability. In this case, there are four classes. The accuracy achieved is 68% and the log-loss result is shown in Fig. 4.

The results of Fig. 3, Fig. 4, and Table I indicate the efficacy of the proposed methodology to identify the mental illness disease using posts and related comments. The main consequences of the experimental results are as follows:

a) The accuracy of training model gradually decreases when the number of training examples is increased, such as training accuracy with 10000 training examples is higher than 15000 examples with minor differences.

b) Cross-validation accuracy gradually increases with the increase in the number of examples. Such as, cross-validation accuracy with 10000 training examples is higher than 5000 examples with minor differences.

c) Class wise classification performance of XGBoost in terms of F-measure remains better. For example, in terms of F-measure, the better performance XGBoost for OCD (0.58), Autism (0.72), PTSD (0.63) and Schizophrenia (0.70) indicate the efficacy of the proposed methodology for the identification mental disorders.

B. *Response to RQ-2.*

In order to respond RQ-2, the performance of the proposed methodology (i.e. XGBoost) is compared with other ML classifiers. According to literature, it has been identified that for the text categorization approach SVM and NB are most accurate and commonly employed classifiers for predicting mental illness. Thus, the performance of XGBoost is compared with SVM and NB. The average performance of NB and SVM is shown in Fig. 5 and Fig. 6, respectively. The results of Fig. 5 and Fig. 6 indicates the performance (Training Score) of SVM which remains better than the performance of NB. For example, the performance of NB (i.e. Accuracy = 0.69) and SVM (i.e. Accuracy = 0.74) with 25000 training examples. However, its performance is less than the performance of XGBoost (i.e. Accuracy = 0.82).

Class wise performance of NB and SVM is shown in Table II and Table III.

The results of Table II and Table III indicates that the class wise performance of NB, and class wise performance of SVM. Such as, the performance of SVM (i.e. F-measure = 0.71) is better than NB (i.e. F-measure = 0.70). However, the performance of SVM is less than the performance of XGBoost (i.e. F-measure = 0.72). Table IV shows comparative assessment of XGBoost, NB and SVM in terms of F-measure.

The result in Table IV indicates the best performance of the proposed model XGBoost in terms of F-measures. The results show that the performance of XGBoost for the classification of OCD comments (F-measure = 0.58), Autism (F-measure = 0.72), PTSD (F-measure = 0.63) and Schizophrenia (F-measure = 0.70) remain better than the performance of NB and SVM.

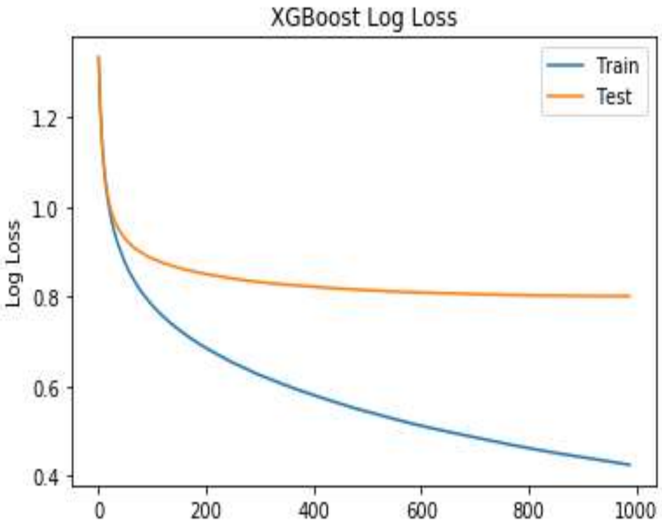


Fig. 4. Log-Loss of XGBoost.

TABLE I. PERFORMANCE EVALUATION OF XGBOOST

Class	Performance Measures		
	Precision	Recall	F-measures
OCD	0.87	0.44	0.58
Autism	0.69	0.76	0.72
PTSD	0.76	0.54	0.63
Schizophrenia	0.60	0.84	0.70

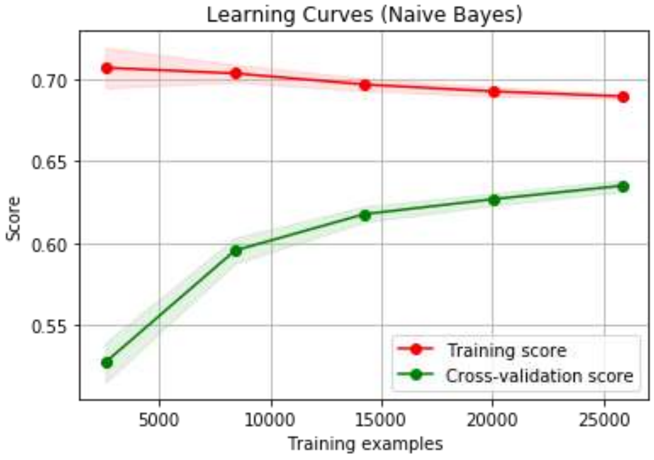


Fig. 5. Learning Curve of NB.

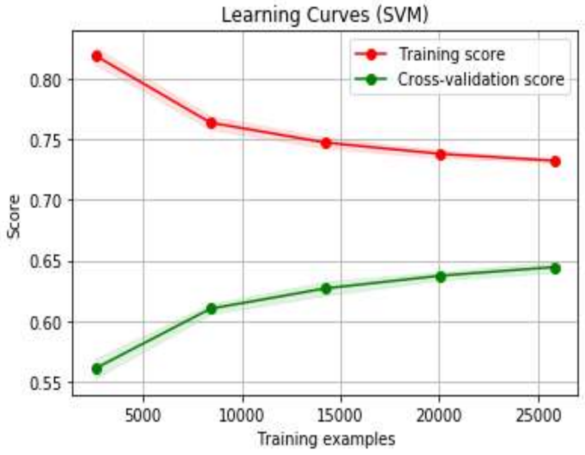


Fig. 6. Learning Curve of SVM.

TABLE II. PERFORMANCE EVALUATION OF NB

Class	Performance Measures		
	Precision	Recall	F-measures
OCD	0.93	0.31	0.47
Autism	0.65	0.76	0.70
PTSD	0.70	0.49	0.57
Schizophrenia	0.52	0.74	0.61

TABLE III. PERFORMANCE EVALUATION OF SVM

Class	Performance Measures		
	Precision	Recall	F-measures
OCD	0.84	0.43	0.57
Autism	0.67	0.75	0.71
PTSD	0.69	0.52	0.60
Schizophrenia	0.55	0.74	0.63

TABLE IV. COMPARISON OF PROPOSED APPROACH WITH STATE OF THE ART CLASSIFIERS

Mental Illness Class	Classifiers Performance		
	NB	SVM	XGBoost
OCD	0.47	0.57	**0.58**
Autism	0.70	0.71	**0.72**
PTSD	0.57	0.60	**0.63**
Schizophrenia	0.61	0.63	**0.70**

V. THREATS TO VALIDITY

In this research work, some treats of using large number of posts and relevant comments are considered that can affect the prediction techniques. The first threat is related to use the limited number of datasets. In this paper, only one dataset is used that is constructed through 32330 comments on 4000 posts. The inclusion of new posts and comments might affect the efficacy the proposed methodology. The second threat is related to comparison of the proposed methodology with the classifiers namely SVM and NB. The finding might be altered when Random Forrest and other widely known outperformed classifiers are reported. Finally, an effectiveness of the proposed methodology is reported by considering the posts and related comments of only four mental illness diseases. The reported effectiveness of proposed methodology might be altered in the existence of posts of more diseases (i.e. class labels in this case).

VI. CONCLUSION

In this paper, a novel ML methodology is proposed to classify the patient's mental illness on the basis of their posts (along with their relevant comments) shared on the well-known social media platform "Reddit". The proposed methodology is exploit by leveraging the capabilities of widely-used classifier namely "XGBoost" for accurate classification of data into four mental disorder classes (Schizophrenia, Autism, OCD and PTSD). The experimental results indicate the effectiveness of the proposed methodology in terms of classification of the mental illness. Though, several mental illness diseases are reported, however, four mental disorder classes are considered, namely schizophrenia, OCD, PSTD, and ASD. A 1000 posts besides related comments are gathered from each of the clinical sub-reddit. Each experiment is performed with k-fold cross validation and used widely known performance measures. The main consequences of the proposed study are as follows: 1) the class-wise performance of XGBosst indicate the effectiveness of the proposed methodology in terms of identifying the mental illness with respect to posts and related comments, Such as the highest F-measure (i.e. 076 for the Autism class and 0.70 for Schizophrenia class); 2) as compared to NB, the XGBoost has made 18.96%, 2.7%, 9.5% and 12.8% improvement in the classification decision for OCD, Autism, PTSD and Schizophrenia, respectively; 3) similarly, as compared to SVM, the XGBoost has made 1.7%, 1.3%, 4.7% and 10% improvement in the classification decision for OCD, Autism, PTSD and Schizophrenia respectively; and 4) the performance of XGBoost remain more significantly better than NB as compared to SVM. In the future work, more classes (i.e. posts of other mental illnesses) can be included to compute the efficacy of the proposed methodology.

REFERENCES

[1] T. B. Üstün, "The global burden of mental disorders," Am. J. Public Health, vol. 89, no. 9, pp. 1315–1318, 1999.

[2] R. C. Kessler et al., "Screening for serious mental illness in the general population," Arch. Gen. Psychiatry, vol. 60, no. 2, pp. 184–189, 2003.

[3] [3] C. Chang et al., "All-cause mortality among people with serious mental illness (SMI), substance use disorders , and depressive disorders in southeast London : a cohort study," 2010.

[4] M. Ernst, J. L. Gowin, C. Gaillard, R. T. Philips, and C. Grillon, "brain sciences Sketching the Power of Machine Learning to Decrypt a Neural Systems Model of Behavior," pp. 1–17, 2019.

[5] M. Srividya, S. Mohanavalli, and N. Bhalaji, "Behavioral Modeling for Mental Health using Machine Learning Algorithms," 2018.

[6] R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, and H. Wang, "Depression detection from social network data using machine learning techniques," Heal. Inf. Sci. Syst., vol. 6, no. 1, pp. 1–12, 2018.

[7] S. F. Behaviors, A. A. Augustine, S. Vazire, D. Ph, N. Holtzman, and S. Gaddis, "Manifestations of Personality in Online Social Networks :," vol. 14, no. 9, 2011.

[8] A. G. Reece and C. M. Danforth, "Instagram photos reveal predictive markers of depression," EPJ Data Sci., 2017.

[9] H. Zhuang et al., "Multimodal classification of drug-naïve first-episode schizophrenia combining anatomical, diffusion and resting state functional resonance imaging," Neurosci. Lett., vol. 705, no. April, pp. 87–93, 2019.

[10] B. Torabi Moghadam et al., "Analyzing DNA methylation patterns in subjects diagnosed with schizophrenia using machine learning methods," J. Psychiatr. Res., vol. 114, no. March, pp. 41–47, 2019.

[11] M. Shim, H. J. Hwang, D. W. Kim, S. H. Lee, and C. H. Im, "Machine-learning-based diagnosis of schizophrenia using combined sensor-level and source-level EEG features," Schizophr. Res., vol. 176, no. 2–3, pp. 314–319, 2016.

[12] H. G. Schnack, "Improving individual predictions: Machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases)," Schizophr. Res., 2017.

[13] K. I. Karstoft, A. Statnikov, S. B. Andersen, T. Madsen, and I. R. Galatzer-Levy, "Early identification of posttraumatic stress following military deployment: Application of machine learning methods to a prospective study of Danish soldiers," J. Affect. Disord., vol. 184, pp. 170–175, 2015.

[14] K. I. Karstoft et al., "Bridging a translational gap: Using machine learning to improve the prediction of PTSD," BMC Psychiatry, vol. 15, no. 1, pp. 1–7, 2015.

[15] G. N. Saxe, S. Ma, J. Ren, and C. Aliferis, "Machine learning methods to predict child posttraumatic stress: A proof of concept study," BMC Psychiatry, vol. 17, no. 1, pp. 1–13, 2017.

[16] J. W. Kanen, "Computational modelling reveals contrasting effects on reinforcement learning and cognitive flexibility in stimulant use disorder and obsessive-compulsive disorder : remediating effects of dopaminergic D2 / 3 receptor agents," 2019.

[17] N. Reggente et al., "Multivariate resting-state functional connectivity predicts response to cognitive behavioral therapy in obsessive–compulsive disorder," Proc. Natl. Acad. Sci., vol. 115, no. 9, pp. 2222–2227, 2018.

[18] F. Lenhard et al., "Prediction of outcome in internet-delivered cognitive behaviour therapy for paediatric obsessive-compulsive disorder: A machine learning approach," Int. J. Methods Psychiatr. Res., vol. 27, no. 1, pp. 1–11, 2018.

[19] J. F. Cohn et al., "Automated Affect Detection in Deep Brain Stimulation for Obsessive-Compulsive Disorder," pp. 40–44, 2018.

[20] S. Levy, M. Duda, N. Haber, and D. P. Wall, "Sparsifying machine learning models identify stable subsets of predictive features for behavioral detection of autism," Mol. Autism, vol. 8, no. 1, pp. 1–17, 2017.

[21] Q. Tariq, J. Daniels, J. N. Schwartz, P. Washington, H. Kalantarian, and D. P. Wall, "Mobile detection of autism through machine learning on home video: A development and prospective validation study," PLoS Med., vol. 15, no. 11, pp. 1–20, 2018.

[22] A. Crippa et al., "Use of Machine Learning to Identify Children with Autism and Their Motor Abnormalities," J. Autism Dev. Disord., vol. 45, no. 7, pp. 2146–2156, 2015.

[23] A. Mahmood and P. Srinivasan, "Twitter bots and gender detection using Tf-idf notebook for PAN at CLEF 2019," CEUR Workshop Proc., vol. 2380, 2019.

You Aren't Alone: Building Arabic Online Supporting Communities using Recommender System

Monirah Alajlan[1], Nouf Alsuhaymi[2], Sara Alnasser[3], Abeer Almohaidib[4]
Nouf Bin Slimah[5], Madawi Alruwaished[6], Najla Alosaimi[7]
Information Systems Department
King Saud University, Riyadh, Saudi Arabia

Abstract—**People are now digitally connected, making the world a single large community. This remarkable benefit has solved many communication issues. For instance, people who go through difficult times and lack the emotional support required to overcome these crises can now join an online support group. For many years, such people had to travel to a predetermined location in a predetermined time to join a support group. Today, with the increasing availability of digital services, these groups can now meet online. For these reasons, this paper presents '*You aren't alone*' mobile application, an interactive mobile-based application designed for Arab people who need psychological support. This application will help in enriching the Arabic content in the field of social support and will help in building supporting communities by peering users to the appropriate support group, anonymously without the fear of judgment. The application will enhance the peering process through a recommender system that reads the user's Twitter timeline and classifies the tweets as belonging to one of the available support groups.**

Keywords—Emotional support; recommender system; classification; support group

I. Introduction

Over the years, technology has changed very rapidly resulting in the emergence of social media which has changed the way we communicate with one another. Currently, there is an extensive variety of social networking sites that create an environment where users can reach the maximum number of people.

Research has shown that the importance of social support is as significant as healthcare support. Recent studies have found that the availability and quality of emotional support have an essential role in prevention or even recovery of illnesses. In fact, the loss or even a lack of social support has been linked to various illnesses and conditions [1]. Recent research has focused on emphasizing the importance of social support networks in the area of communication technologies [2]. These include understating the importance of social support networks and their psychological effects and how we might apply such knowledge to design and develop relevant interventions.

On a local scale, the social support issue is alarming and the need for support groups is rising due to the increase in the prevalence of psychological issues which need support, such as patients with psychological disorders, diabetes mellitus, hypertension, bronchial asthma, and others. However, few studies have been conducted in Saudi Arabia to determine the prevalence of psychological issues among Saudi people. A study conducted by AlKahtani [3], measured the prevalence rate of psychological issues in a random sample of Saudi adults, aged from 15 to 65 years, which was found to be 18.2%. Additionally, a study conducted by Al-Sughayr and Ferwana [4] covered a random sample of high school students and showed that the prevalence rate of psychological issues was 48%. Therefore, this problem requires proper intervention in the form of social support. Unfortunately, to the best of our knowledge, literature review on Arabic applications revealed the fact that there is a lack of applications that provide psychological support and thus, this application will address this issue.

Considering the critical need for social support in Saudi Arabia, 'You aren't alone' application aims to fulfil the following objectives. First, to help and support the country's people and those of the Arab region in general who lack emotional support. Second, to include and support people who undergo similar issues regardless of their psychological nature. Such issues might be as simple as the struggles of new mothers, the confusion of fresh graduates, and the disappointment of unemployed graduates. Third, the application aims to build a network of people who have similar issues, in order to share and view thoughts, stories, and experiences. Moreover, they can create their own virtual reading rooms and ask questions, and life coaches and experts may volunteer to help and answer questions. Furthermore, the application aims to enhance the process of peering the users with the appropriate supporting group through Twitter data mining.

The application can suggest a support group for the user by linking their account to Twitter. As nowadays Twitter is considered as an expressive platform where people's thoughts and emotions can be reflected by their tweets. By using Twitter's API to stream user's tweets and then applying topic modeling techniques to classify tweets into their relevant support group. Twitter is an online microblogging tool that disseminates more than 400 million messages per day [5]. As the tweets allow to gain an insight into the online public behavior, it represents an important data source to conduct textual analysis. Twitter is considered more effective than other social media platforms for analysis because the tweets are publicly available and easily accessible via APIs [6].

This paper is structured as follows. Section II illustrates the literature review that discusses related studies and applications.

A comprehensive comparison against similar systems is shown in Section III. Section IV presents the application development process used in the project. Section V discusses the limitations and future work, and Section VI concludes the paper.

II. Literature Review

This section begins with a literature review to present previous studies that discussed topics related to the concept of mental health. Next, a comparison of similar applications based on differing features is presented. Finally, we present a discussion that summarizes the results and outcomes.

A. Related Studies

Life can be hard for many people. Unfortunately, a lot of people who suffer from issues that are triggered by life's events such as loss of loved ones, family-related issues, isolation, and many more end up with psychological distress. A study in [7] revealed that individuals with low connectedness exhibit more dysfunctional interpersonal behaviors which in turn contribute to more psychological distress. Similarly, individuals with high connectedness exhibit more appropriate interpersonal behaviors which in turn contribute to less psychological distress [7]. In the same vein, a study in [8] investigated the effectiveness of joining support groups for patients with contact dermatitis, reported that 77.8% of the patients noticed a positive difference in the way they understand and cope with this illness. This positive effect was also reported in [9], where the COVID-19 pandemic has caused enormous stress on people working in the health care sector. The study concluded that joining a virtual support group was indeed helpful during the crisis.

Another problem with social support is that some people find it embarrassing to reveal their identity in a setting where they express their deepest emotions and pain. This might be because people are usually afraid of other people's judgment. Research indicates that even experts often judge others and that the diagnostic judgments made by counselors are often influenced by their preferences for different types of counseling problems and their biases in the use of diagnostic classification systems. Studies have proven that the final number of counseling interviews is affected by the type of the client's issue and certain of their descriptive characteristics [10].

By conceptualizing social networking platforms as new social domains, online social support could supplement the mental health benefits derived from in-person social support [11]. Research suggests that people who are more introverted or socially anxious might be more likely than others to derive benefit from social media. Particularly, there is an absence of supportive apps that can help people who need emotional support and reduce their sense of loneliness in the Arab world [9]. To the best of our knowledge, our search in Android and Apple markets yielded no apps dedicated to this issue.

Twitter has increasingly been used by individuals to express their thoughts, and feelings in the form of short text messages. In their 2014 study [12], Hasan et al. concluded that classifying short texts according to finer-grained classes of emotions provides rich and informative data about the emotional states of individuals. In addition, these data can be used by healthcare professionals for early detection of psychological disorders such as anxiety or depression.

B. Related Applications

- Sanvello

Sanvello is a website and a mobile app for both iOS and android [13]. It aims to help people suffering from stress, anxiety, and depression and it offers several tools to address mental health issues. These tools are based on cognitive behavior therapy, mindfulness meditation, relaxation, and mood/health tracking. Users can go by a nickname when they share their stories, advice, and more with Sanvello's peer support community via a specified topic discussion or a chat group. The application provides special features such as when it is first launched, the user will get a message telling him "take a moment to pause and think of a place that relaxes you". Subsequently, the user will choose a theme, set goals, and build a daily habit of self-care.

- Lyf

Lyf is a social media app where users share their feelings and experiences by posting, connecting, and chatting with other users who share the same concern [14]. In the "Feed", users can post their thoughts in a timeline. It illustrates the "stream" in which users share their journeys as private, public, or anonymous to share stories sequentially and to keep updating on their story. Users can also chat with other people with the same concern or with "live support" which is an online counselling platform for users who seek professional advice and guidance.

- We are close

It is an official android and iOS application introduced by the Saudi Ministry of Health [15]. This application allows members to read topics and articles related to mental health with the aim to educate people about mental health issues. Interestingly, members can consult with specialists by filling a form with information such as the age, gender, the problem for consulting, medications they use, and any other health problems. The application also has a section for voice messages to communicate directly with the health specialists. Additionally, there is a section of interaction in which the member can add an article or a picture or write a motivational story. There are books that tackle mental health issues that have been made available to users on the application. The benefit of this application is that users do not have to register first to use the application; they can take a look and use some features as guest.

- ReachOut

ReachOut is to support network for patients and care givers. Essentially, it is an iOS and android application but has a blog dedicated to inspiring, educating, and connecting readers with real-world health stories and articles. It is a peer support app categorized into seven categories for mental health, cancer, diabetes, chronic pain, etc. Users can choose more than one category. The primary interaction is through a timeline where users share their thoughts and others can "give a hug" or write a comment. Users also can communicate with each other through messages which facilitate private communication supporting and sharing more feelings. To use the application, first one has to create an account and then manage their profile by filling the name, age, status, and gender; subsequently, one will be able to use the application.

III. Comparison with Related Applications

Comparing similar applications in a table demonstrates the differences between the applications easily. Table I illustrates the five applications including our application and compares them in terms of multiple features including functionality, platforms, and the content. It can be seen that 'You aren't alone' has several of the important features that exist in similar applications, such as the timeline and anonymity. On the other hand, it provides features that are not supported in any other applications like suggestion based on tweets.

TABLE I. Comparison with Similar Systems

Features\App	**Pacifica**	**Lyf**	**We are close**	**ReachOut**	**You Aren't Alone**
Timeline	✓	✓	✓	✓	✓
Anonymous	✓	✓	✓	✗	✓
Likes	✓	✗	✓	✓	✓
Join specific support group	✓	✗	✗	✓	✓
Create a support group	✗	✗	✗	✗	✓
Participation of Therapist and life coaches	✗	✓	✓	✗	✓
Share articles	✗	✗	✓	✓	✓
Suggest group based on twitter	✗	✗	✗	✗	✓
Question and Answers	✗	✗	✓	✗	✓
Subscription fees for more features	✓	✓	✓	✓	✗
Website	✓	✗	✗	✓	✗
Mobile Application	✓	✓	✓	✓	✓
Arabic	✗	✗	✓	✗	✓

IV. Application Development

A. Information Gathering

1) *Interview*: The interview is one of the most important techniques for information gathering; it helps collect a richer source of information on behavior, opinions, knowledge, and preferences. Therefore, three experts were selected to gain insights from their experiences.

Several interviews were conducted with specialists in the field of psychology. A profound physician stated that psychological support is a psychological need for each human being. He explains the concept of Group Psychotherapy as where people share their experience, symptoms and experience in treatment. It's a codified type of psychological treatment.

Effective communication reduces the symptoms of psychological illness. Also, effective communication has a chemical effect also, it changes the proportion of chemicals in the body that affect the receptors of those chemicals in the body.

Furthermore, a college professor confirmed for us the importance of psychological support and considered it the basis for maintaining good mental health. When asked about the importance of sharing similar experiences, she said that research in developed countries has revealed that people should engage in psychosocial sessions because sharing feelings with others helps to alleviate the feelings but locking it up can turn it from a problem situation into a state of depression, nervousness or tension and then social isolation which may affects the person negatively. In the same vein, a lecturer stressed the importance of supporting the individual first and the society second. An individual is the basic building block of society and by supporting him we might create a healthy society with awareness and freedom. Moreover, sharing feelings with others might facilitate social interaction and emotional support. However, sometimes people have difficulty sharing feelings because they may expect no reaction to their feelings or may have been shy since childhood. Unfortunately, not sharing tragedies or personal issues might lead to several behavioral problems such as social shyness and lack of self-confidence, which affects productivity.

2) *Questionnaire*: Based on a study covering a sample of 937 people aged 14 and above, we found out that native Arabic speakers are facing issues with the availability of resources in social and emotional support as more than 50% of the people questioned agreed that there is a lack of Arabic content for psychological and social support platforms. For instance, 53.7% of the sample have faced difficult times and experiences and sought support with no luck, while others wanted to contact life coaches and therapists but never got the opportunity. We also found out that a significant percentage of those who were questioned are feeling a lot of social pressure as 54% of the respondents wanted to remain anonymous while expressing their true feelings, and more than 50% of the respondents felt that they fear judgments, blame, and misunderstanding when sharing personal stories. On the other hand, we found that there are some improvement mechanisms

that motivate people to connect and share their feelings. For instance, most respondents stated that they prefer talking to people with similar personal experiences as such people understand them which makes them feel better. Interestingly, most respondents stated that social support improves confidence and sense of self-fulfillment, and 77% of respondents indicated that volunteering and community service improves their psychological wellbeing. Another 84% were willing to provide help to those in need for emotional and psychological support. We also found that people care about sharing valuable news and research papers as 90% feel better when friends share articles and news that matters to them. When asked about their feelings about developing an application that provides psychological health care and support groups in Arabic, more than 96% of respondents chose the option “great idea”. Based on the aforementioned results and due to the lack of available resources in social support and dearth of Arabic content, we aim to provide a social support platform in Arabic to help Arab communities that need emotional support share their experiences and relieve all the ambiguous emotions they might be feeling by connecting them with a group of people with similar experiences along with the supervision of therapists and life coaches who would love to join voluntarily.

B. *System Implementation*

You aren’t alone system is composed of three main components. The first one is the application program which is developed using Xcode IDE platform that supports iOS application development with Swift programming language. The Interface Builder within Xcode is used to design the user interface. The second component is Firebase Real-time database which is responsible for storing retrieving and manipulating system's data. Firebase real-time database is a cloud-hosted database which is a collection of objects stored in a tree-like hierarchical structure. Fig. 1 illustrates the database collections.

The third component is the data mining component, where a user’s tweets are mined in order to produce suitable support groups. Twitter is an online microblogging tool that disseminates more than 400 million messages per day [5]. As the tweets allow to gain an insight into the online public behavior, it represents an important data source to conduct textual analysis. The “Suggest for me” function goes through several phases as shown in Fig. 2.

1) Data collection: Twitter Kit is a native SDK to include Twitter content in mobile apps [16]. It provides many features like displaying Tweets, authorizing Twitter users, and working with the Twitter API. First, using “*Log in With Twitter*” feature provided by Twitter kit to enables app users to authenticate with Twitter. Then building a query to retrieve random tweets from the public timeline. A dataset of over 500 tweets was collected and stored in a .csv file.

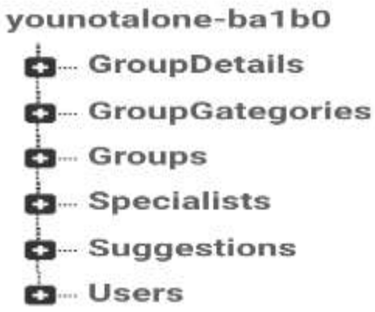

younotalone-ba1b0
GroupDetails
GroupGategories
Groups
Specialists
Suggestions
Users

Fig. 1. Firebase Database Collection.

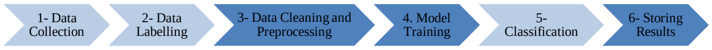


Fig. 2. ‘Suggest for Me’ Phases.

2) Dataset labelling: After collecting tweets from the public timeline, manual labelling by the team was done, in order to classify each tweet as related to one of the ten support groups available such as depression and diabetes.

3) Data cleaning and pre-processing: Data go through two steps cleaning and pre-processing. In cleaning, the punctuation, stop words, and URLs are removed because they carry little meaning. Whereas in pre-processing, the text is normalized to scale the data into fewer ranges.

In the case of suggesting support groups based on the users’ tweets, it was recognized that there are some words that carry more meaning than other words. For instance, the world ‘depression’ is more valuable than the world ‘having’. In order to eliminate the worthless words before it takes valuable processing time, NLTK was use. The NLTK corpus library detect and extract the commonly used stop words. For example, in this tweet: “I need to improve myself to save my relationship." After running the code using NLTK to detect and remove the stop words the output was: ['need', 'improve', 'myself', 'save', 'relationship'].

4) Model training: In this step, one of the Natural Language Processing techniques, the topic modelling is used. Topic modelling is a process to automatically identify topics present in a text object and to derive hidden patterns exhibited by a text corpus [17]. An example of topic modelling algorithm is Latent Dirichlet Allocation (LDA), it is used to discover themes in text and classify text into those predetermined themes.

Using LDA topic modeling to analyze the tweets and their related support groups, a dictionary of keywords for each support group is created. This is accomplished through calculating the probabilities of words and deciding which of them are closely related to the topic of the support group. The dictionary is divided into 10 categories, each of them represents a group and contains words that reflect the emotion related to the group. Table II shows sample terms from the created dictionary.

TABLE II. SAMPLE TERMS FROM THE DATASET

Stress	**Addiction**	**Personal Development**	**Depression**	**Health**	**Obsessive compulsive Disorder**	**Hyperactivity**	**Social Wellbeing**	**Cancer**	**Diabetes**
Overthinking	Addiction	Self esteem	Feel lonely	Diet	OCD	ADHD	Fight	Chemotherapy	Diabetes
Stressed out	Smoke	Self identify	Sad	Health	Obsessive compulsive disorder	Hyperactivity	Hate	Leukemia	TID
Stress	Drug	Insecurity	Worried	Weight	Compulsions	Hyper	Rude	Breast cancer	T2D
anxious	cigarettes	confident	depressed	food	obsessions	Extreme hyperactivity	problems	Lung cancer	insulin

5) Classification: In this phase, the system mainly works with an end-point API to preform NLP classification. Firstly, using LDA topic classifier to analyze the user profile, specifically the last 10 tweets on his timeline. In order to return group suggestions to the user given the last ten tweets, the model calculates and tests the probabilities for each class resembling the group topic and compares them to the stored key words from the initial training dataset. Then, it returns a list of support groups with the highest probabilities based on the tweets' text. In this case, the results contain the suggested group, confidence and the reason.

6) Storing results: Once the analysis gets the appropriate groups for the user, it's then automatically added to the user joined groups. As it is a one-time process there is no need to store the analysis result nor the user twitter information.

C. System Design

For 'You aren't alone' application, the most suitable architecture is a combination of Model View Presenter (MVP) and client server architectures as shown in Fig. 3. Client-server architecture can be defined as a software architecture made up of both the client and server, whereby the clients always send requests while the server responds to the requests sent [18].

Since the application, as a client, requires accessing several servers such as Firebase and Twitter API in order to establish a connection and send requests, the client-server approach was a logical solution [19]. Furthermore, MVP allows for decoupling code in presenter level and by that enabling reuse of the code interface of the application without the need of using third party tools.

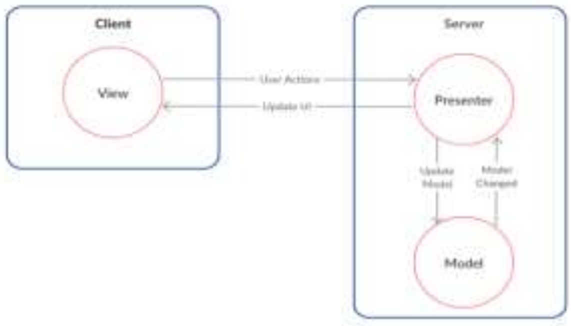


Fig. 3. You Aren't Alone Architecture.

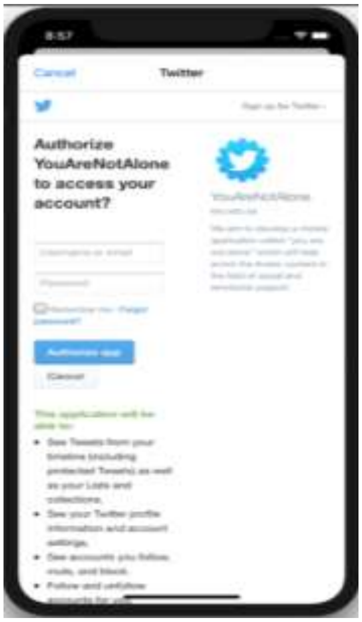
D. User Interface

This section presents a sample of screens in You aren't alone to illustrate the user interface design (Fig. 4 to 7).

E. System Evaluation and Testing

1) Acceptance testing: User Acceptance Testing (UAT) plays an important role in validating the requirement by target users to prevent losses after releasing the system. UAT was conducted with users from different ages and different concerns to make sure the system provides a variety of support groups.

Two stages of testing were conducted. The first stage is to calculate the time each user spent on each task. The second stage is a survey on the same participants with 7 questions regarding system's functionality, clearance, user-friendliness, interfaces. Table III displays a sample of the user acceptance testing. Based on the results the users took much longer time in posting a thought and asking questions. It takes longer for the users to type than to view as well as expressing their feelings might need a longer time.



Cancel Twitter

Sign up for Twitter ›

Authorize YouAreNotAlone to access your account?

YouAreNotAlone

We aim to develop a mobile application called "you are not alone" which will [illegible] in the field of social and emotional support.

Username or email

Password

□ Remember me · Forgot password?

Authorize app

Cancel

This application will be able to:

- See Tweets from your timeline (including protected Tweets) as well as your Lists and collections.
- See your Twitter profile information and account settings.
- See accounts you follow, mute, and block.
- Follow and unfollow accounts for you

Fig. 4. Suggest for Me Interface.

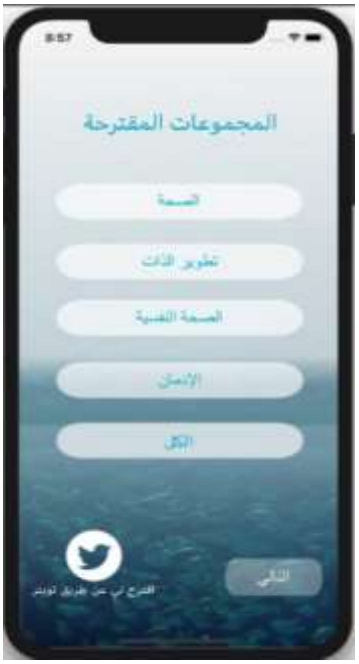


Fig. 5. Suggested Groups Interface.

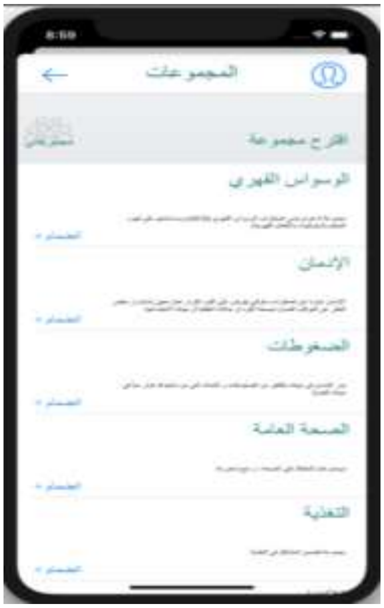
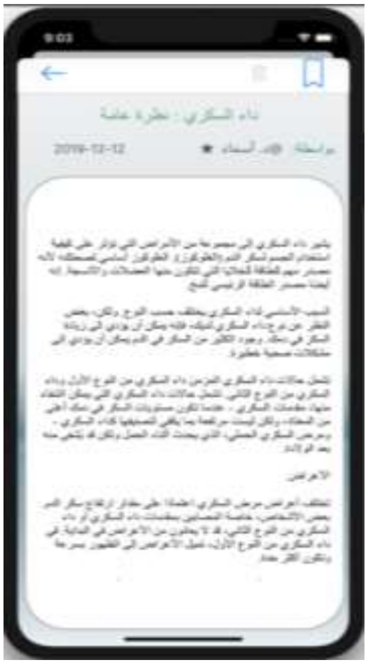


Fig. 6. Support Groups to be Joined Manually.



داء السكري : نظرة عامة

بواسطة @د. أسماء ★ 2019-12-12

يشير داء السكري إلى مجموعة من الأمراض التي تؤثر على كيفية استخدام الجسم لسكر الدم (الجلوكوز). الجلوكوز أساسي لصحتك؛ لأنه مصدر مهم للطاقة للخلايا التي تكون منها العضلات والأنسجة. إنه أيضا مصدر الطاقة الرئيسي للمخ.

السبب الأساسي لداء السكري يختلف حسب النوع. ولكن، بغض النظر عن نوع داء السكري لديك، فإنه يمكن أن يؤدي إلى زيادة السكر في دمك. وجود الكثير من السكر في الدم يمكن أن يؤدي إلى مشكلات صحية خطيرة.

تشمل حالات داء السكري المزمن داء السكري من النوع الأول وداء السكري من النوع الثاني. تشمل حالات داء السكري التي يمكن الشفاء منها: مقدمات السكري - عندما تكون مستويات السكر في دمك أعلى من المعتاد، ولكن ليست مرتفعة بما يكفي لتصنيفها كداء السكري - ومرض السكري الحملي، الذي يحدث أثناء الحمل ولكن قد يُشفى منه بعد الولادة.

الأعراض

تختلف أعراض مرض السكري اعتمادًا على مقدار ارتفاع سكر الدم. بعض الأشخاص، خاصة المصابين بمقدمات داء السكري أو داء السكري من النوع الثاني، قد لا يعانون من الأعراض في البداية. في داء السكري من النوع الأول، تميل الأعراض إلى الظهور بسرعة وتكون أكثر حدة.

Fig. 7. Articles Interface.

TABLE III. ACCEPTANCE TESTING RESULTS

Tester name	**Haifa Majed (Member)**			
Task	***Number of errors***	***Time needed***	***User's feedback***	***Completion status***
Browsing home page	0	00:00:36 sec	-	Completed.
Linking to twitter.	0	00:00:11 sec	"Great characteristic"	Completed.
Suggest a support group	0	00:00:42 sec	-	Completed.
Browsing support groups.	0	00:00:08 sec	-	Completed.
Posting a thought	0	00:00:46 sec	-	Completed.
Check the timeline	0	00:00:07 sec	"Presented very well"	Completed.
Reply a thought	0	00:00:21 sec	-	Completed.
Share articles	0	00:00:22 sec	-	Completed.
Pin articles	0	00:00:04 sec	-	Completed.
Ask questions	0	00:00:30 sec	-	Completed.
Answer to questions	0	00:00:16 sec	-	Completed.
Browse the whole app	0	00:04:05 min	-	Completed.

V. LIMITATIONS AND FUTURE WORK

One of the limitations of the system would be that it is only available for iOS users and due to this limitation, we intend to extend the scope of the system to support a Web application platform and Android devices. Another limitation is that it only extracts textual features form Twitter using text mining. However, deep emotions are sometimes hidden in the images a user post.

The system currently supports Arabic language only, but we plan to expand our system to include universal languages (ex. English). Another future direction would be to improve the recommendation process by, expanding the scope of the system to generate recommendations of support groups based on user past behaviors rather than the content of users' tweets. Past behavior includes the browsing habits and frequently used website since they give a strong indication of his psychological state and/or mood.

VI. CONCLUSION

In this paper, 'you aren't alone' application, an interactive mobile application dedicated for those who need social support in the form of support groups, was introduced. Although people are becoming more aware about the psychological issues, many problems arise when a person wants to join a support group, such as the lack of confidence and the fear of judgment. The application aims towards advancing the mental and psychological health of people by providing adequate support groups. Despite the numerous applications in this field, 'you aren't alone' is one of a few applications that targets the Arab community and the first to suggest a support group based on a user's tweets. The team hope that 'You aren't alone' becomes a supportive and inviting environment where people struggling in life, find it a safe place to freely express their feelings and engage with others.

ACKNOWLEDGMENT

This research project was supported by a grant from the "Research Center of the Female Scientific and Medical Colleges", Deanship of Scientific Research, King Saud University.

REFERENCES

[1] Marilyn Frank-Stromborg, Sharon J. Olsen. "Instruments for clinical Health-care Research". Dec 2004.

[2] Heather Attig. "Social support and new communication technologies during a life stressor". Nov 2013.

[3] Al-Khathami AD, Ogbeide DO. "Prevalence of mental illness among Saudi adult primary-care patients in Central Saudi Arabia". Saudi Med J., Jun 2002.

[4] Al-Sughayr AM, Ferwana MS." Prevalence of mental disorders among high school students in National Guard Housing, Riyadh, Saudi Arabia". J Fam Community Med, Jan 2012.

[5] A. Kim, J. Murphy, J. Duke, H. Hansen, A. Richards, and J. Allen. "Methodological Considerations in Analyzing Twitter Data". Journal of the National Cancer Institute. Monographs. 2013.

[6] Van den Rul, C. (2019). A Guide to Mining and Analysing Tweets with R. Medium. Available at: https://towardsdatascience.com/a-guide-to-mining-and-analysing-tweets-with-r-2f56818fdd16. [Accessed online on 9/12/2019].

[7] Richard M., Matthew Draper and Sujin Lee. "Social Connectedness, Dysfunctional Interpersonal Behaviors, and Psychological Distress: Testing a Mediator Model", Journal of Counseling Psychology. Vol. 48, No. 3, 310-318, 2011.

[8] L. Voller, S. Kullberg, Y. Liou and S. Hylwa, "Effectiveness of Support Groups for Patients With Contact Dermatitis: A Pilot Study", Dermatitis, vol. 31, no. 6, pp. 383-388, 2020.

[9] Viswanathan, M. Myers and A. Fanous, "Support Groups and Individual Mental Health Care via Video Conferencing for Frontline Clinicians During the COVID-19 Pandemic", Psychosomatics, vol. 61, no. 5, pp. 538-543, 2020.

[10] John B., Richard S., Deborah M. "Counselor Intake Judgments, Client Characteristics, and Number of Sessions at a University Counseling Center", Journal of Counseling Psychology.975, Vol. 22, No. 6, 567-559.

[11] David A., Elizabeth A., Rachel L., Kathryn M., and Tawny Spinelli. "Online Social Support for Young People: Does It Recapitulate In-person Social Support; Can It Help?". Dec 2016.

[12] M. Hasan, E. Rundensteiner, and E. Agu, "EMOTEX: Detecting Emotions in Twitter Messages" 2014 ASE bigdata/socialcom/cybersecurity Conference. Available at: http://web.cs.wpi.edu/~emmanuel/publications/PDFs/C30.pdf.

[13] Sanvello A Place to Feel Better, https://www.sanvello.com/, [Accessed November 29, 2020].

[14] *Lyf*. Lyf App Ltd, 2020 [Online]. Available: https://play.google.com/store/apps/details?id=com.incogo.android&hl=en&gl=US, [Accessed November 21, 2020].

[15] We are close: Your own Clinic at Home, Available: http://ncmh.org.sa/index.php/pages/view/90/14/14, [Accessed November 28, 2020].

[16] Tools and Libraries, Twitter Inc, [online]. Available: https://developer.twitter.com/en/docs/tools-and-libraries.

[17] Beginners Guide to Topic Modeling in Python and Feature Selection. Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/. [Accessed online on 10/12/2019].

[18] H. Oluwatosin. "Client-Server Model", IOSR Journal of Computer Engineering, vol. 16, no. 1, pp. 57-71, 2014.

[19] V. Corneliu, M.Iulian, Octavian. "Model View Presenter Design Pattern", Journal of Computer Science and Control Systems. vol. 3, pp. 173-176, 2010.

Facebook Profile Credibility Detection using Machine and Deep Learning Techniques based on User's Sentiment Response on Status Message

Esraa A. Afify[1]*, Ahmed Sharaf Eldin[2]
Information Systems Department
Faculty of Computers and Artificial Intelligence
Helwan University, Cairo, Egypt

Ayman E. Khedr[3]
Information Systems Department
Faculty of Computers and Information Technology
Future University in Egypt (FUE), Cairo, Egypt

***Abstract*—Recently, the impact of online Social Network sites (SNS) has dramatically changed, and fake accounts became a vital issue that has rapidly evolved. This issue gives rise to how to assess and measure the credibility of User-Generated Content (UGC). This content is used in finding trusted sources of information on SNS like Facebook, Twitter, etc. Consequently, classifying users' profiles and analyzing each user's behavior response based on the content generated became a challenge that must be solved. One of the most significant approaches is Sentiment Analysis (SA) which plays a major role in assessing and detecting the credibility degree of each user account behavior. In this paper, the aim of the study is to measure and predict the user's profile credibility by declaring the correlation degree among the UGC features that affect users' responses to status messages. The proposed models were implemented using six Supervised Machine Learning classifiers, an Unsupervised Machine Learning cluster model, and a Deep Learning Neural Network (NN) model. The research paper presents two experiments to evaluate Facebook profile credibility. At first, we applied a binary classification model to classify profiles into fake or genuine users. Then, we conducted a classification model on genuine users based on the credibility theory by using the Analytical Hierarchical Process (AHP) approach and computed the credibility score for each. Secondly, we selected and analyzed a public Facebook page (CNN public page) and obtained data from it for users' sentiment reactions and responses on statuses Messages relating to different topics on the period (2016/2017). Then, we performed LDA on the status corpus (Topic Modeling algorithm, Latent Dirichlet Allocation) to generate topic vectors. In addition, we performed Principal Component Analysis (PCA) method to visualize and classify each status topic distribution. Afterthought, we produced a status corpus cluster to classify users' behaviors through statuses posted and users' comments. As a conclusion of this study, the first experimental results achieved 95% and 99% accuracy to classify fake/genuine users and incredible/credible accounts, respectively. The second experiment outcome identified the clusters for the status corpus in 10 topic-features distribution and classified users' contents into credible or not according to the final calculated credibility score.**

Keywords*—*Fake profiles detection; credible profiles detection; sentiment analysis; supervised machine learning classifiers; unsupervised machine learning; binary classification; deep learning neural network; evaluation metrics

I. Introduction

Social Networks (SN) became the primary activity in our lives and turned out to be a virtual community [1]. In a real community, people massively exchange their opinions in every aspect of life. Some people could be considered as credible ones, and others are not according to the availability and reliability them. Usually, we accept other opinions according to the activeness behavior for each of them. Applying the same concept to the virtual SN community, people create posts and comments as if they are in real life through a variety of social accounts. Then, they interact with them in which raises the need to detect unreliable contents created in SNs [2].

Facebook and Twitter are Social Network Sites (SNS) that have experienced a dramatic increase in popularity over the last few years. Especially, Fake profiles on Facebook which harm privacy, online bullying, misuse, and trolling, etc. These profiles related to users with false credentials. It could be found through malicious and undesirable activities, causing problems for social network users. Users create fake profiles for social engineering, online representation to slander an individual, advertising, and campaigning for an individual or a group of individuals.

According to the Pew Research Center, Facebook has reached a leading position among the SNSs, with some worldwide active users amounting to over 2.3 billion as of July 2017. The main feature of Facebook and other SNSs is the possibility for users to share self-generated content like texts, pictures, audio, and video with their friends or followers. Users could create or share fake content because of missing approaches used to measure the credibility of the generated or shared content. On public pages of Facebook, users are not allowed to post, but they can only contribute by commenting on the posts. Sometimes users' input is unrelated to the post, for example, the topic of the post and the comment is different, or the comment is spam. Not only the comments of users on the page post are essential for measuring the credibility of the post, but also there are other features like the number of reactions, the number of shares, and Facebook emotions including "angry", "wow", "haha", "love", and "sad" reactions on posts, comments, and even messages,

*Corresponding Author

which could be used for measuring the credibility degree of the generated content.

The researchers found several characteristics and patterns that could be used to identify the credibility degree of user profile and user action/ interaction behavior. Then, they focused on, Sentiment Analysis (SA) which leads to figuring out how people feel about social media. With a sophisticated analysis of how people react to certain topics, we can predict various issues such as campaign success, marketing strategy, product messaging, customer service, and stock market price. As a result, we decided to take advantage of the recent extensions of reactions made by Facebook and do sentiment analysis on how people react differently to different posts. Based on the credibility theory, we used the Analytical Hierarchical Processes (AHP) approach to produce the feature weights to compute the credibility score for each user profile. After that, we analyzed users' sentiment analysis and performed LDA on the status corpus (Topic Modeling algorithm, Latent Dirichlet Allocation) to cluster topic-features distribution and Principal Component Analysis (PCA) method to visualize and classify each status topic distribution to compute a credibility score. Machine learning techniques contribute efficiently to detect semantic relations [3] in general and frauds [4] in specific. According to the revolution in Artificial Intelligence (AI), [5][6][7], we found that Machine Learning (ML) and Deep Learning (DL) are leading in research to predict the models' performance. For this reason in this research paper, we followed the ML and DL pipeline and performed two models for detecting the credible score of the users' profile and the content shared by them on social networks by discovering new patterns and characteristics for each user's profile. The first model is a binary classification model that automatically detects the fake and genuine profiles on Facebook. This model implies six supervised machine learning classifiers like Support Vector Machine, Random Forest, Decision Tree, K-Nearest Neighbor, Logistic Regression, Naïve Bayes, and a deep learning Neural Network model to classify the profiles into fake or genuine. The second model is a clustering model that detects credible and non-credible profiles according to user behaviors using the sentiment analysis generate on each profile. This model applied using the K-Means unsupervised machine learning clustering. Different performance analysis approaches conducted to evaluate both experiments such as plotting the Learning Curves (LC), calculating the "Area Under the Curve" (AUC) of "Receiver Characteristic Operator" (ROC), illustrating the ROC/AUC Curves, computing the Confusion Metrics (CM), and generating classification reports to summarize results for each applied classifier.

Research paper organization. This paper is organized as follows: Section II briefly discusses the related works to the research study. Section III presents the research methodologies. Section IV describes the proposed methodology. Section V identifies the results and discussion of the experiments. And Section VI provides the research study conclusion.

II. Literature Review

Extracting semantic relations has been successfully applied. As found in, Sultan et al. (2012) [8], semantic relations exchange is performed for information sources' collaboration. This approach would support different sources including Facebook for detection. Another research in a different direction, as in Sharaf Eldin et al. (2015) [9], focused on detecting the appropriate technique for the type of data as successful techniques determination is one of the key success factors.

Focusing on Facebook sources concerning credibility detection on Facebook, the most recent researches are: Lê et al. (2019) [10], proposed a ranking scheme for fake Facebook user accounts detection. The model includes both feature-based approaches and graph-based approaches by utilizing the SVM and SybilWalk algorithm.

Smruthi et al. (2019) [11], used a hybrid model based on machine learning and skin detection algorithms to detect the existence of fake accounts on Facebook. The model result achieved 80% accuracy by utilizing the supervised machine learning algorithms.

Gupta et al. (2017) [1], attempted to detect fake accounts on Facebook based on user profile activities and interactions. The model result achieved 79% accuracy by applying the most supervised machine learning algorithms.

Wani et al., (2016) [12], presented a novel approach to predict fake profiles on Facebook. The model was trained using supervised machine learning algorithms. The theoretical machine learning model has been proposed to classify the user profiles into fake and genuine.

Saikaew et al., (2015) [2], developed a system for measuring credibility on Facebook information. At first, the authors proposed a FB credibility evaluator. Secondly, they developed a chrome extension to evaluate the credibility of each post. Based on the usage analysis of their FB credibility chrome extension, about 81% of users responded agree with suggested credibility automatically computed by the proposed system.

III. Research Methodologies

A. Machine Learning (ML): Overview

ML is the main branch of computer science that, provides computers with the capacity to learn without being programmed. It begins with data extracting knowledge. In *ML*, a dataset of observations, called vectors, comprises several variables called features or attributes [13].

In the next sections, we will discuss the two main categories of machine learning, which are supervised learning and unsupervised learning. In this paper, we used supervised learning for the first experiment and unsupervised learning for the second one.

B. Supervised Learning: Methods

Supervised learning, also known as predictive modeling, is the process of making predictions using pre-labeled data. As shown in Fig. 1, it takes input datasets with output labels. This data called ‘training data’ that include a set of training examples [14]. A subclass of supervised learning problems is binary classification, where there are only two labels for class features as a fake class or genuine class.

In our first proposed model, the dataset is a series of fake and genuine users’ profiles, our supervised task is to predict whether each user account is fake or genuine. First, we train a classifier using the existing label. Labeled data with the desired output is called ‘model training’ because the model is learning the relationship between the attributes (features) of the data and the desired output value (target). These features include the number of friends, number of followers, statuses, gender, and language, and so on. Second, we make predictions for the new data for which we do not know the true outcome. For example, when a new user account created, we want our trained model to accurately predict whether the user account is fake or genuine without a human examination. The best-case scenario will allow the classifier to correctly set the class labels for unseen cases. This is supervised learning because there is a specific outcome we are trying to predict, in our work namely, fake, or genuine users.

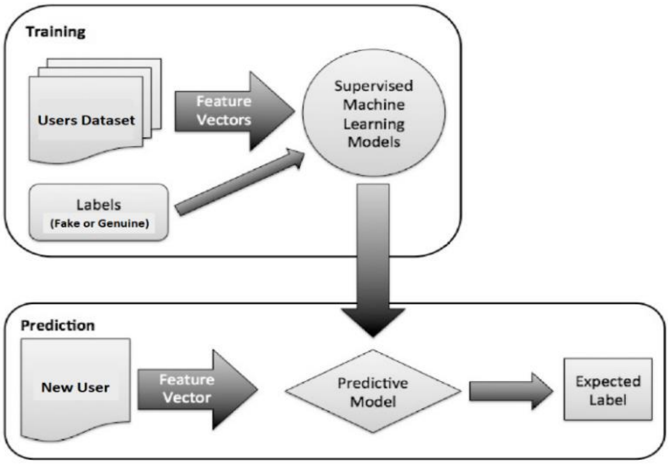


Fig. 1. Supervised Machine Learning.

In the next section, we will discuss briefly the six classification models that have been selected to implement the research work on this paper.

C. Classification Models: Brief

Classification, known as an instance of Supervised Machine Learning, is a method of setting to which class does a new observation belongs, based on training the machine with an existing data containing observations, in which class is predefined. The algorithms which implement Classification are called as Classifiers, there are many types of classifiers available, as follows:

1) Decision Tree (DT) Classifier applies a hierarchical structure, each internal node denotes a test on an attribute. It breaks down the dataset to build the model. Each node classifies an output value of a test and every leaf or terminal node holds a class label. This classifier splits the tree in the target variable that is most dominant, after calculating the entropy and gain scores [15] [16].

$$\text{Entropy (s)} = \sum_{\mathrm{i=1}}^{\mathrm{c}} - \mathrm{p_i} \log_2 \mathrm{p_i} \tag{1}$$

$$\text{Gain (S, A)} = \underbrace{\text{Entropy(S)}}_{\text{Original entropy of S}} - \underbrace{\sum_{v \in values(A)} \frac{|Sv|}{|S|} \,.\, Entropy\ (Sv)}_{\text{relative entropy of S}} \tag{2}$$

2) Random Forest (RF) Classifier based on ensemble learning. It combines multiple decision trees to form a strong classifier [17]. In each decision tree, we pick a random sample from the training set, then choose random features at each node of the tree. After that, we split the tree at the best split among the selected features. In the binary classification case, the result is the percentage of trees that give a majority voting score.

$$\mathrm{RFfi_i} = \frac{\sum_j norm\ fi_{ij}}{\sum_{j \in all\ features, k\ \in all\ trees} normfi_{jk}} \tag{3}$$

3) Logistic Regression (LR) Classifier uses a *sigmoid function* [7] [15], as shown in *Fig. 2*. It maps predicted observations to estimate probabilities between 0 and 1 or True/False.

$$f(x) = \sigma(x) = \frac{1}{1+e^{-x}} \tag{4}$$

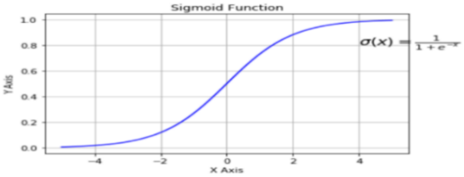


Fig. 2. Sigmoid Function.

4) K-Nearest Neighbor (KNN) Classifier based on similarity measures or distance functions. It uses a *K* value to get the nearest neighbor class, then performs a majority voting. The KNN calculates the numerical values using *distance formulas* [7] [18] [19].

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{x} (x_t - y_t)^2} \tag{5}$$

$$\text{Manhattan distance} = \sum_{i=1}^{k} |x_i - y_i| \tag{6}$$

$$\text{Minkowski distance} = \left(\sum_{i=1}^{k} (|x_i - y_i|)^q\right)^{1/q} \tag{7}$$

5) Naïve Bayes (NB) Classifier based on Bayes' Theorem and conditional probability. It uses Bayes' formula to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of the prediction [7] [15].

$$p(c|x) = \frac{p(X|C)\, p(c)}{p(x)} \tag{8}$$

$$p(c|x) = p(x_1|c) \times p(x_2|c) \times \ldots \times p(x_n|c) \times p(c) \tag{9}$$

- $P(c|x)$ is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- $P(x|c)$ is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

6) Support Vector Machine (SVM) Classifier plots each observation as a point in n-dimensional space (n refers to features). After that, it finds the optimal hyper-plane by maximizing the margins between classes, as shown in Fig. 3, [20][21].

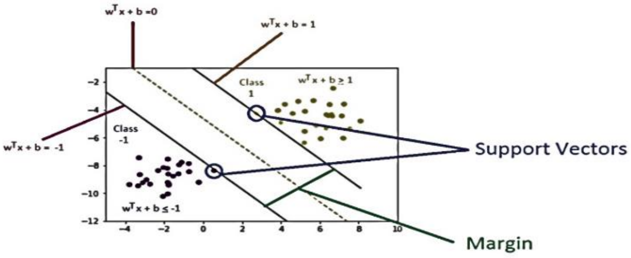


Fig. 3. Support Vector Machine.

D. *Artificial Neural Network Deep Learning: Overview*

Deep learning, known as, a subset of machine learning that does a similar function, but there are many layers, every layer provides a different performance to the data it feeds on, as shown in Fig. 4, for example. The name Artificial Neural Network (ANN) came from, functioning as an inspiration, or as it works as the function of the neural networks present in the human brain [5] [6] [22]. Recently, deep learning is the evolution of machine learning, which performs as a neural network that vest machines to produce accurate decisions without humans interfering.

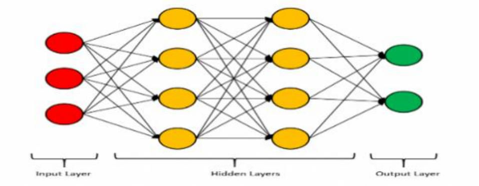


Fig. 4. ANN with Two Hidden Layers.

E. *Unsupervised Learning: Methods*

Unsupervised learning, also known as a data-driven model, is the process of identifying clusters using unlabeled data. It takes input dataset only where patterns or structures are found as hidden features among the dataset. This training dataset is a collection of observation examples without a specific desired outcome. Clustering is a typical example of unsupervised learning that finds visual classifications that match hypotheses. The purpose of clustering is to bring similarities, regardless of the data class. Therefore, a clustering algorithm usually, needs to know how to calculate the similarity, then start to run.

K-Means Clustering is a clustering algorithm that combines the n of observations into k clusters that aggregated with each other, according to specific similarities [6] [23], as shown in Fig. 5.

It works according to three steps, as follows:

- Initialization – K initial “means” (centroids) generated randomly.
- Assignment – K clusters created by associating each observation with the nearest centroid.
- Update – The centroid of the clusters becomes the new mean.

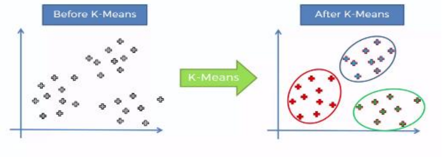


Fig. 5. K-Means Clustering.

F. *Evaluation Curves and Metrics for a Classification Model*

We can evaluate the classification model with different curves and metrics, such as Learning Curves, AUC-ROC Curves, Confusion Matrix, Accuracy Score, Precision Score, Recall Score, F1 Score, and Specificity [24].

Learning Curve: used to plot each classification model. These plots used for visualizing the observations with the metric performance. Line of learning plotted the y-axis over the experience of the x-axis to model the training set performance against the set as a function of the training set size. In a learning curve, a good fit is clarified by a training and validation loss that decreased to a point of stability with a small gap between both final loss outputs.

ROC/AUC Curve: used to plot the 'true positive rate' illustrated on the y-axis versus the 'false positive rate' which illustrated on the x-axis for the whole potential classification thresholds.

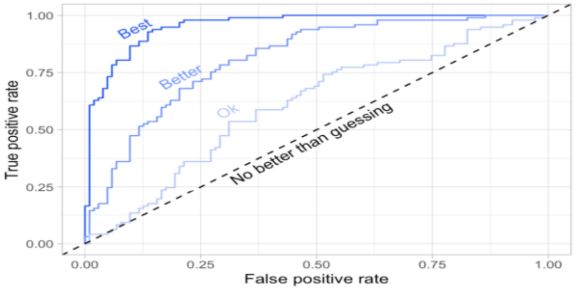


Fig. 6. ROC / AUC Curve.

As shown in Fig. 6, to utilize this terminology, 'sensitivity' defined on the y-axis and 1 minus specificity on the x-axis for every classification threshold from zero to one. Also, the dashed line in the graph is the baseline state the random guesses where the 'true positive rate' increases linearly with the 'false positive rate', and its AUC is 0.5; the blue line is the ROC plot of the model, and its AUC is less than 1. In a perfect case, the ‘true positive’ samples have a probability 1, so that the ROC starts at the point with 100% 'true positive' and 0 'false positives'. The AUC of such a perfect curve is 1. A line that is diagonal from the lower left corner to the upper right corner represents a random guess. The higher the line is in the upper left-hand corner, the better.

Confusion Matrix: is a table with four different combinations of predicted and actual values. It illustrates all the observations in the testing set. In other words, it summarizes predicted outcomes and true outcomes for testing, as presented in Table I.

TABLE I. CONFUSION MATRIX

		Predicted		

		Negative	Positive	TN = True Negative
				FP = False Positive
				FN = False Negative
				TP = True Positive

Actual	Negative	TN	FP	
	Positive	FN	TP	

- TN is the false sample, which is predicted to be false by the model.
- FP is the false sample, which is predicted to be true by the model.
- FN is the positive sample, which is predicted to be false by the model.
- TP is the positive sample, which is predicted to be true by the model.

The calculation formulas of *FPR* and *TPR* are as follows:

$$FPR = \frac{FP}{TN+FP}, TPR = \frac{TP}{TP+FN} \tag{10}$$

$$Accuracy = \frac{|TP|+|TN|}{|TP|+|TN|+|FP|+|FN|} \tag{11}$$

$$Precision = \frac{|TP|}{|TP|+|FP|} \tag{12}$$

$$Recall = \frac{|TP|}{|TP|+|FN|} \tag{13}$$

$$F1 = 2.\frac{Precision.Recall}{Precision+Recall} \tag{14}$$

$$Specifity = \frac{|TN|}{|TN|+|FP|} \tag{15}$$

$$AUC = \frac{\sum(n_0 + n_1 + 1 - r_i) - n_0(n_0+1)/2}{n_0 n_1} \tag{16}$$

IV. PROPOSED METHODOLOGY

In this research paper, we aim to propose a set of minimum features that can detect credible users' profiles and sentiment responses with the highest accuracy into two models. To do that, we followed the general machine learning and deep learning pipeline step-by-step, as shown in Fig. 7.

A. Data Acquisition: Datasets

Two different datasets had been used to implement our proposed models. Firstly, we applied the classification model on a public dataset that consists of 2818 fake and genuine users' profiles with 34 features, but after applying the correlation for them, we extracted 7 features that affect the detection method. Secondly, the cluster model had been applied on a CNN public Facebook page. This data related to various users' sentiment responses at 10 different topics distribution on status messages during the period (2016/2017). The dataset consists of 9282 status messages, with 14 features. The experiments implementation was deployed by python code on Google Colab Notebooks and applied using Machine Learning models with the help of the Scikit-learn libraries. Keras with TensorFlow used for Deep Learning model.

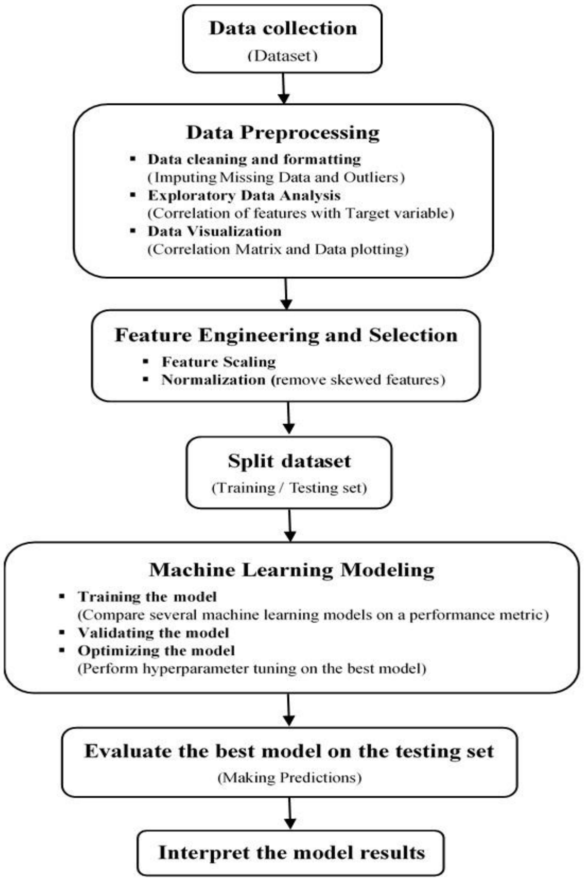

Data collection
(Dataset)
Data Preprocessing
Data cleaning and formatting
(Imputing Missing Data and Outliers)
Exploratory Data Analysis
(Correlation of features with Target variable)
Data Visualization
(Correlation Matrix and Data plotting)
Feature Engineering and Selection
Feature Scaling
Normalization (remove skewed features)
Split dataset
(Training / Testing set)
Machine Learning Modeling
Training the model
(Compare several machine learning models on a performance metric)
Validating the model
Optimizing the model
(Perform hyperparameter tuning on the best model)
Evaluate the best model on the testing set
(Making Predictions)
Interpret the model results

Fig. 7. ML and DL Pipeline.

B. Data Pre-Processing

1) Data cleaning using outliers detection: The Tukey's boxplot method [25], as shown in Fig. 8, considered to be one of the most frequently used methods for finding outliers uses the interquartile range with boxplot to filter out exceptionally large or ridiculously small numbers whether a distribution is skewed and whether there are potential unusual observations in the dataset.

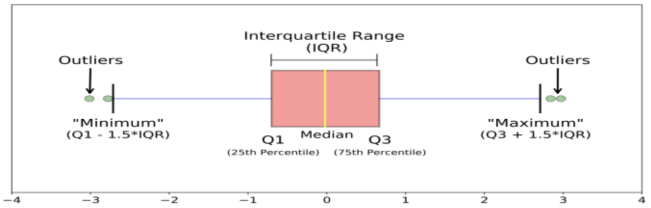


Fig. 8. Tukey's Method (Box Whisker).

Algorithm 1: Pseudo-code for Outliers Detection using the Tukey Method

Input: Dataset

Output: Suspected Outliers data points

Procedure: First quartile 25% (Q1), Third quartile 75% (Q3), Interquartile Range 50% (IQR)

```
for data values $d_i$ in the Training dataset do
    Arrange $d_i \rightarrow$ Q1 and $d_i \rightarrow$ Q3
    Compute IQR = Q3 – Q1
    Compute the outlier boundaries formulas, as follows:
        Lower Outlier Boundary $l_i$ = Q1 – 1.5 (IQR)
        Upper Outlier Boundary $u_i$ = Q3 + 1.5 (IQR)
    if $d_i < l_i$ or $d_i > u_i$ then
        return Outliers
    end if
end for
```

We have detected and eliminated the outliers from the Facebook CNN public page dataset following Algorithm 1 proved above. We visualized the boxplots and removed all outliers in each user sentiment response to achieve the best results during the experiment testing, as shown in Fig. 9 and 10, respectively.

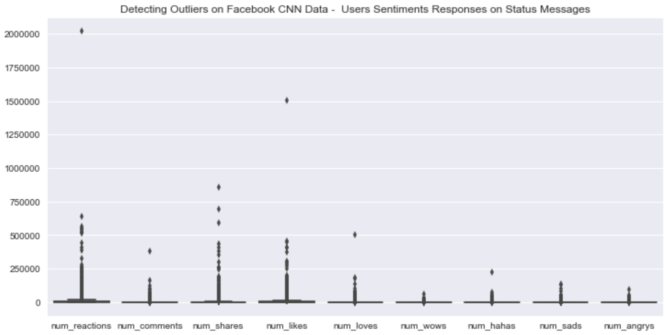


Fig. 9. Detecting Outliers on Facebook CNN Page - Users' Responses.

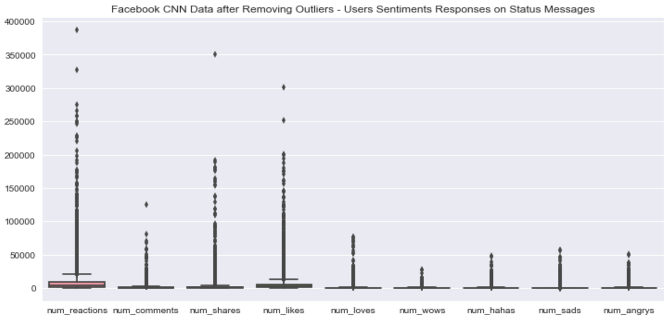


Fig. 10. Eliminating Outliers from Facebook CNN Data - Users' Responses.

2) Data analysis: elbow method and silhouette score method: A fundamental step for any unsupervised algorithm is to determine the optimal k number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods used to determine this optimal value of k, as shown in Algorithm 2 and Fig. 11.

Algorithm 2: Pseudo-code for Elbow & Silhouette Method

Input: Data $X = \{\mathbf{X_1}, \ldots, \mathbf{X}_n\}$, the order k, **MAX** number of allowed iterations

Output: A partition P = $\{C_1, \ldots, C_\mathbf{k}\}$

```
$t = 0$, $\mathbf{P} = \varnothing$
Randomly initialize $\mu_i$, $i = 1, \ldots, k$
loop
    $t += 1$
    Assignment Step: assign each sample $\mathbf{x}_j$ to the cluster with the
    nearest representative
    $c_i^{(t)} = \{\mathbf{X_j} : d(\mathbf{X_j}, \mu_i) \leq d(\mathbf{X_j}, \mu_h)$ for all $h = 1, \ldots, k\}$
    Update Step: update the representatives
    $\mu_i^{(t+1)} = \frac{1}{|c_1^{(t)}|} \sum_{xj} \in c_i^{xj}$
    Update the partition with the modified clusters:
    $P^t = \{c_1^{(t)}, \ldots, c_k^{(t)}\}$
    if $t \geq$ MAX OR $P^t = P^{t-1}$ then
        return $P^t$
    end if
end loop
```

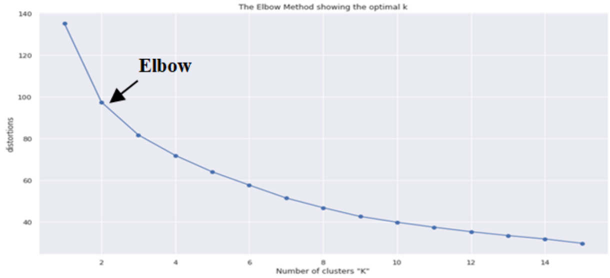


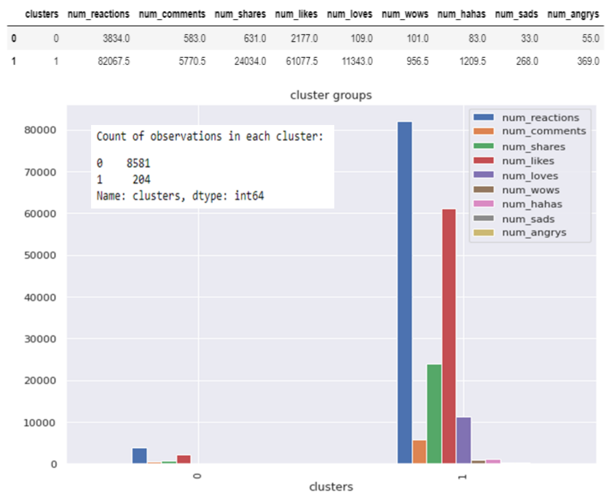
Fig. 11. Selecting the Number of Clusters k using the "Elbow Method".

In this paper the researchers had used the 'elbow method' to specify the number of clusters k that the algorithm must find to define the user's profiles groups numbers. This curve has the shape of an arm, the "elbow" found at k=2 in this model. Where the distortions illustrated on the y-axis, then dropped very quickly as the k increased up to 2, then it decreased much more slowly as the k increased more, which illustrated on the x-axis.

In Fig. 12, the observations divided into two groups of users:

- ***Credible users:*** *'cluster 0'*, this group of users are not extensively using Facebook a lot and only use it for surfing. The reaction count is only 48 on posts and comments 3. And they did not share any posts and only react 47 '*like*' on posts.
- ***Non-credible users:*** *'cluster 1'*, this group of users are extensively using Facebook. They react to 82067 posts and comments 57770. And they share posts and use the other reacts on posts.

The Silhouette considered being a better method to choose the optimal number of clusters k to be formulated from the data. This method measures the similarity of a data instance within a cluster comparing with another cluster. Then computes the score for each data instance and calculate the formula for the Silhouette coefficient as shown in Fig. 13.



	clusters	num_reactions	num_comments	num_shares	num_likes	num_loves	num_wows	num_hahas	num_sads	num_angrys
0	0	3834.0	583.0	631.0	2177.0	109.0	101.0	83.0	33.0	55.0
1	1	82067.5	5770.5	24034.0	61077.5	11343.0	956.5	1209.5	268.0	369.0

```
Count of observations in each cluster:

0    8581
1     204
Name: clusters, dtype: int64
```

Fig. 12. Visualizing the Clusters Groups.

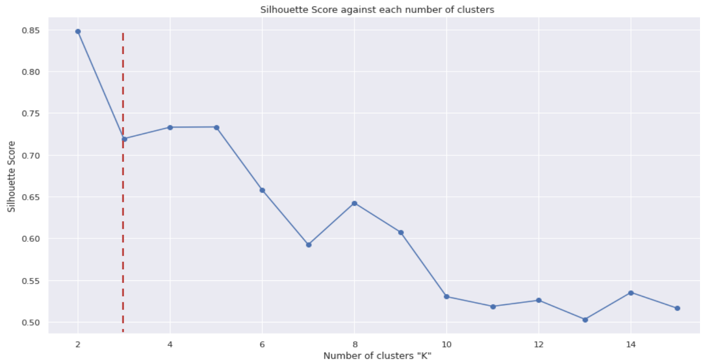
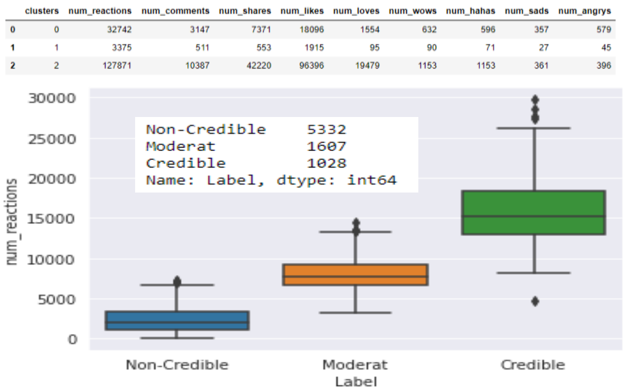


Fig. 13. Selecting the Number of Clusters k using the "Silhouette Score".

From the pivoted data frame shown in Fig. 14, we can see that there are three groups of Facebook users:

- ***Group 0:*** This is the group of people who, according to the provided dataset, happen to use Facebook quite a lot. But they are the kind of people who usually give people the 'like' react mostly.
- ***Group 1:*** Which indicates that the user of this group might not use Facebook a lot or use it only for surfing. Their number of reactions are around 3375 and comments only 511. They do not share a lot of posts. And mostly they use 'like' react on posts.
- ***Group 2:*** This group also shows that people use Facebook a lot. These people tend to comment and share the posts a lot. They also tend to use other reacts on posts besides the 'like' react.

3) Data visualization: correlation coefficient matrix: Visualizing the Correlation Matrices after dealing with null values, dropping unnecessary features, and eliminating outliers from the dataset, as shown in Fig. 15 and 16.

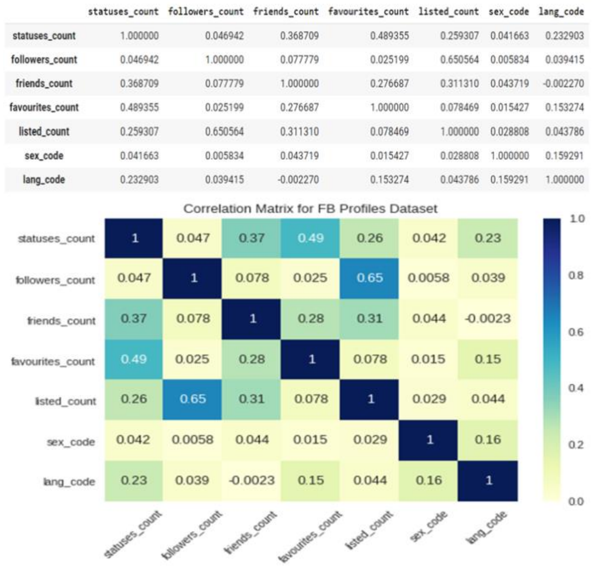


	clusters	num_reactions	num_comments	num_shares	num_likes	num_loves	num_wows	num_hahas	num_sads	num_angrys
0	0	32742	3147	7371	18096	1554	632	596	357	579
1	1	3375	511	553	1915	95	90	71	27	45
2	2	127871	10387	42220	96396	19479	1153	1153	361	396

```
Non-Credible     5332
Moderat          1607
Credible         1028
Name: Label, dtype: int64
```

Fig. 14. Visualizing the Clusters Groups from the Pivoted Data Frame.

Model 1

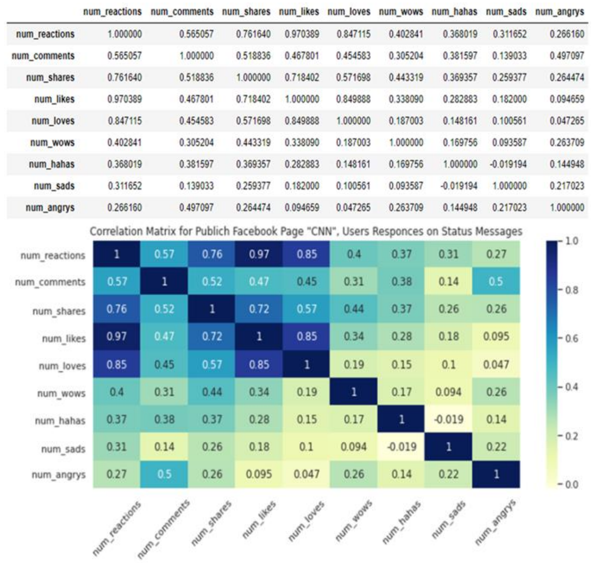


	statuses_count	followers_count	friends_count	favourites_count	listed_count	sex_code	lang_code
statuses_count	1.000000	0.046942	0.368709	0.489355	0.259307	0.041663	0.232903
followers_count	0.046942	1.000000	0.077779	0.025199	0.650564	0.005834	0.039415
friends_count	0.368709	0.077779	1.000000	0.276687	0.311310	0.043719	-0.002270
favourites_count	0.489355	0.025199	0.276687	1.000000	0.078469	0.015427	0.153274
listed_count	0.259307	0.650564	0.311310	0.078469	1.000000	0.028808	0.043786
sex_code	0.041663	0.005834	0.043719	0.015427	0.028808	1.000000	0.159291
lang_code	0.232903	0.039415	-0.002270	0.153274	0.043786	0.159291	1.000000

Correlation Matrix for FB Profiles Dataset
statuses_count 1 0.047 0.37 0.49 0.26 0.042 0.23
followers_count 0.047 1 0.078 0.025 0.65 0.0058 0.039
friends_count 0.37 0.078 1 0.28 0.31 0.044 -0.0023
favourites_count 0.49 0.025 0.28 1 0.078 0.015 0.15
listed_count 0.26 0.65 0.31 0.078 1 0.029 0.044
sex_code 0.042 0.0058 0.044 0.015 0.029 1 0.16
lang_code 0.23 0.039 -0.0023 0.15 0.044 0.16 1
statuses_count
followers_count
friends_count
favourites_count
listed_count
sex_code
lang_code
1.0
0.8
0.6
0.4
0.2
0.0

Fig. 15. Facebook user Profile Correlation Matrix.

Model 2:



	num_reactions	num_comments	num_shares	num_likes	num_loves	num_wows	num_hahas	num_sads	num_angrys
num_reactions	1.000000	0.565057	0.761640	0.970389	0.847115	0.402841	0.368019	0.311652	0.266160
num_comments	0.565057	1.000000	0.518836	0.467801	0.454583	0.305204	0.381597	0.139033	0.497097
num_shares	0.761640	0.518836	1.000000	0.718402	0.571698	0.443319	0.369357	0.259377	0.264474
num_likes	0.970389	0.467801	0.718402	1.000000	0.849888	0.338090	0.282883	0.182000	0.094659
num_loves	0.847115	0.454583	0.571698	0.849888	1.000000	0.187003	0.148161	0.100561	0.047265
num_wows	0.402841	0.305204	0.443319	0.338090	0.187003	1.000000	0.169756	0.093587	0.263709
num_hahas	0.368019	0.381597	0.369357	0.282883	0.148161	0.169756	1.000000	-0.019194	0.144948
num_sads	0.311652	0.139033	0.259377	0.182000	0.100561	0.093587	-0.019194	1.000000	0.217023
num_angrys	0.266160	0.497097	0.264474	0.094659	0.047265	0.263709	0.144948	0.217023	1.000000

Correlation Matrix for Publich Facebook Page "CNN", Users Responces on Status Messages
num_reactions 1 0.57 0.76 0.97 0.85 0.4 0.37 0.31 0.27
num_comments 0.57 1 0.52 0.47 0.45 0.31 0.38 0.14 0.5
num_shares 0.76 0.52 1 0.72 0.57 0.44 0.37 0.26 0.26
num_likes 0.97 0.47 0.72 1 0.85 0.34 0.28 0.18 0.095
num_loves 0.85 0.45 0.57 0.85 1 0.19 0.15 0.1 0.047
num_wows 0.4 0.31 0.44 0.34 0.19 1 0.17 0.094 0.26
num_hahas 0.37 0.38 0.37 0.28 0.15 0.17 1 -0.019 0.14
num_sads 0.31 0.14 0.26 0.18 0.1 0.094 -0.019 1 0.22
num_angrys 0.27 0.5 0.26 0.095 0.047 0.26 0.14 0.22 1
num_reactions
num_comments
num_shares
num_likes
num_loves
num_wows
num_hahas
num_sads
num_angrys
1.0
0.8
0.6
0.4
0.2
0.0

Fig. 16. Facebook CNN Page - Users' Sentiments Correlation Matrix.

C. Feature Engineering and Selection: Steps

Each feature in the dataset has a degree of importance to represent the data very well. In consequence, the feature selection step is needed like a filter, wrapper, and embedded method. One of the Topic Modeling methods is Latent Dirichlet Allocation (LDA). LDA is used to classify text in a collection of group documents by topics, as described in the following steps and shown in Algorithm 3.

The step-by-step approach for LDA with classifiers explained below:

- Read the data which comprises a combination of genuine and fake users.
- Pre-process the data to filter out status messages in genuine users' case.
- Prepare every user data by concatenating entire posts for user.
- Apply the LDA algorithm on posts after concatenation to generate topics.
- Generates user or post probabilities of *n* topics.
- Evaluate the loss and Goss metrics for every user or post.
- Use the vectors set of features for training classifiers.
- Classify the feature vector into train/test set then train with models.
- Report and compute accuracy, recall, f-score and precision of the algorithm.

Algorithm 3: Pseudo-code for LDA

```
1: Choose distribution of topic
2: $\theta a$ ~Dirichlet (α) where $a \in \{1, …., X\}$ and Dirichlet (α) is the
   Dirichlet distribution for α parameter
3: For every word $W_{ab}$ in the document where $b \in \{1, … ... N_a\}$
4: Select a particular topic $z_{ab}$ ~ $Multi$ $(\theta a)$ where multi ( ) is
   a multinomial
5: Select a word $W_{ab} \sim \beta Z_{ab}$
where w indicates words, Z indicates topic vector and $\beta$ is a $K$ x $V$
matrix of word probability for every term (column) and every topic
(row) and $\beta_{ab} = P(W_{a\,=}\,1|Z^{a}=1)$
```

In addition, dimensionality reduction considered a type of feature selection applied for significant large features. One of the most well-known dimensionality reduction methods called Principal component analysis (PCA). PCA method is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components [26]. PCA is mostly used as a tool in exploratory data analysis and for making predictive models. It is often used to visualize genetic distance and relatedness between populations. PCA is either done in the following five steps as shown in Algorithm 4.

Algorithm 4: Pseudo-code for PCA

```
1: Compute the mean feature vector
   $\mu = \frac{1}{p} \sum_{k=1}^{p} x_k$, where, $x_k$ is a pattern
2: Find the covariance matrix
$C = \frac{1}{p} \sum_{k=1}^{p} \{x_k - \mu\}^T$ where, T represents matrix transposition
3: Compute Eigen values $\lambda_i$ and Eigen vectors $v_i$ of covariance
   matrix
$Cv_i = \lambda_i v_i$ $(i = 1, 2, 3, .....q)$, $q$ = number of features
4: Estimating high-valued Eigen vectors
   (i) Arrange all the Eigen values ($\lambda_i$) in descending order
   (ii) Choose a threshold value, $\theta$
   (iii) Number of high-valued $\lambda_i$ can be chosen to satisfy the
         relationship
         $[\sum_{i=1}^{s} \lambda_i]\ [\sum_{i=1}^{p} \lambda_i]^{-1} \geq \theta$, where, s = number of high valued $\lambda_i$
         chosen
   (iv) Select Eigen vectors corresponding to selected high valued $\lambda_i$
5: Extract low dimensional feature vectors (principal components)
   from raw feature matrix.
   $P - V^{T}x$, where, $V$ is the matrix of principal components and $x$ is
   the feature matrix
```

The first proposed model consists of various steps:

- Determines the main account features that influence a correct detection of fake profiles,
- Apply and compare different classification algorithm,
- Illustrate and compute the evaluation curves and metrics for each classifier, and.
- Compute the credibility score for the genuine users' accounts by using the AHP approach, as shown in algorithm 5.

The second model also consists of various steps:

- Select and acquire data from a public Facebook page for user sentiment analysis,
- Determine the main features that influence users' profile behaviors, through status message and users' responses,
- Perform LDA topic modeling algorithm on status corpus and generate topic vectors,
- Assign for each status a most relevant topic label based on highest probability,
- Perform PCA to visualize topic distribution and correlation matrix,
- Analyze and visualize users' responses on each topic,
- Apply a K-Mean clustering algorithm to cluster status corpus using topic-features,
- Plot likelihood/inertia for each K-number of clusters for each method, and

- Compute the credibility score for users' responses on status corpus by using the AHP approach.

Algorithm 5: Pseudo-code for AHP Approach

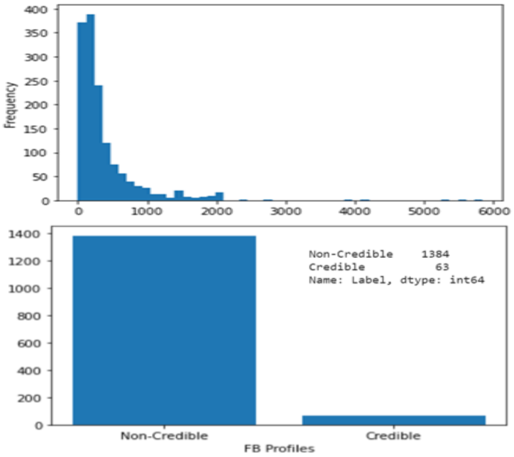
```
Input: Dataset
Output: Alternatives Ranking
Procedure:
for data values $d_i$ in the Training dataset do
    - Construct the AHP Hierarchy for evaluation:
          Level 1 → define a decision goal
          Level 2 → set the criterion
          Level 3 → distribute the alterative
    - Calculate the Pairwise Comparison Matrix (Matrix A)
          $A=\begin{bmatrix} a11 & a12 \cdots & a1n \\ a21 & a22 \cdots & a2n \\ \vdots & \vdots \cdots & \vdots \\ an1 & an2 \cdots & ann \end{bmatrix}$, where: $a_{ij} = 1/a_{ji}$ (I,j = 1,2,..,n)
    - Calculate Normalized principal Eigen Vector of Matrix A 'w' (Priority Vector Matrix
          $e^T = (1, 1, \ldots..,1) \rightarrow W = \lim_{k\rightarrow\infty} \frac{A^K.e}{e^T.A^K.e}$
          $Aw = \lambda max\ w \rightarrow \lambda max \geq n$
          $\lambda max = \frac{\sum^{ajwj-n}}{w1}$
          A={aij} with aij=1/ aij
          Where:
          A → pair wise comparison
          W → normalized weight vector
          λmax → maximum eigen value of matrix A
          aij → numerical comparison between the values i and j
    - Calculate the weights and testing the consistency for each level
    - Calculate Consistency Ratio
          $Calculate\ Consistency\ Ratio\ (CR) = \frac{Consistency\ Index\ (CI)}{Random\ Consistency\ Index\ (RI)} \rightarrow CR = \frac{CI}{RI},$
          Where:
          $CI = \frac{\lambda\ max - n}{n-1}$
    if Matrix Consistence, CR ≤ 0.10 then
        Get the priorities of all selection criteria
        Get the rank of each alternative with respect to the selection criteria
        return Get the overall rank of the alternatives
    end if
end for
```

D. Credibility Detection Method: Formulas

In the proposed model, we proposed a credibility formula for both genuine profiles and status messages. This formula contains several parameters each of these parameters multiplied with a specific weight define according to the correlation coefficient matrix. These weights computed according to the Analytical Hierarchical Process (AHP) approach, which depends on the credibility theory. Applying this equation will lead us to rank users' accounts each according to the credibility ranking. Consequently, we can predict the degree of trust and credibility of Facebook user profiles, as shown in Fig. 17 and 18.

1) Facebook Profile Credibility Formula:

Profile Credibility Degree = Statues count * 0.33
+ Followers count * 0.23
+ Friends count * 0.16
+ Favorites count * 0.13
+ List count * 0.08
+ Gender code * 0.04
+ Language code * 0.02 (17)



```
Non-Credible    1384
Credible          63
Name: Label, dtype: int64
```

Fig. 17. Count of Credible users' Profiles Plot.

2) Facebook Status Message Credibility Formula:

Status Credibility Degree = $num_{reactions}$ * 0.26
+ $num_{comments}$ * 0.189
+ num_{shares} * 0.12
+ num_{likes} * 0.17
+ num_{loves} * 0.107
+ num_{wows} * 0.075
+ num_{hahas} * 0.046
+ num_{sads} * 0.027
+ num_{angrys} * 0.014 (18)

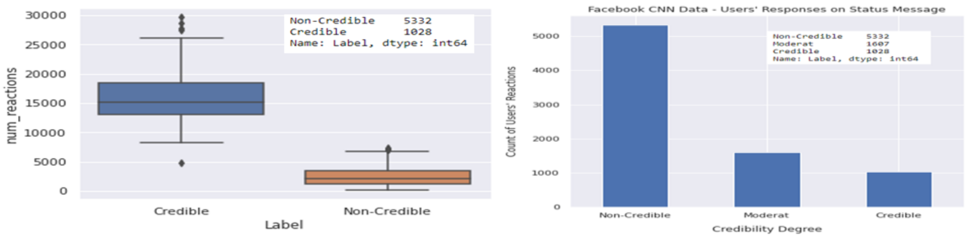
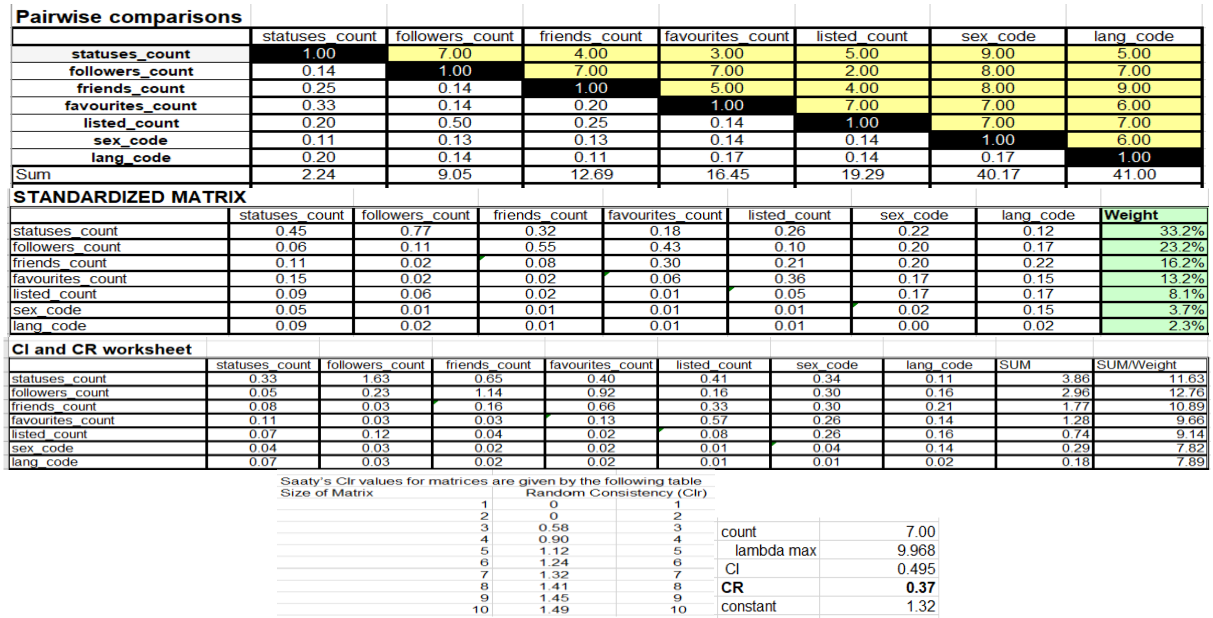


Fig. 18. Count of Credible users' Responses Plot.

E. AHP: Calculation

In the following Fig. 19 and 20, we will present how weights computed according to the Analytical Hierarchical Process (AHP) approach in details.



Pairwise comparisons

	statuses_count	followers_count	friends_count	favourites_count	listed_count	sex_code	lang_code
statuses_count	1.00	7.00	4.00	3.00	5.00	9.00	5.00
followers_count	0.14	1.00	7.00	7.00	2.00	8.00	7.00
friends_count	0.25	0.14	1.00	5.00	4.00	8.00	9.00
favourites_count	0.33	0.14	0.20	1.00	7.00	7.00	6.00
listed_count	0.20	0.50	0.25	0.14	1.00	7.00	7.00
sex_code	0.11	0.13	0.13	0.14	0.14	1.00	6.00
lang_code	0.20	0.14	0.11	0.17	0.14	0.17	1.00
Sum	2.24	9.05	12.69	16.45	19.29	40.17	41.00

STANDARDIZED MATRIX

	statuses_count	followers_count	friends_count	favourites_count	listed_count	sex_code	lang_code	Weight
statuses_count	0.45	0.77	0.32	0.18	0.26	0.22	0.12	33.2%
followers_count	0.06	0.11	0.55	0.43	0.10	0.20	0.17	23.2%
friends_count	0.11	0.02	0.08	0.30	0.21	0.20	0.22	16.2%
favourites_count	0.15	0.02	0.02	0.06	0.36	0.17	0.15	13.2%
listed_count	0.09	0.06	0.02	0.01	0.05	0.17	0.17	8.1%
sex_code	0.05	0.01	0.01	0.01	0.01	0.02	0.15	3.7%
lang_code	0.09	0.02	0.01	0.01	0.01	0.00	0.02	2.3%

CI and CR worksheet

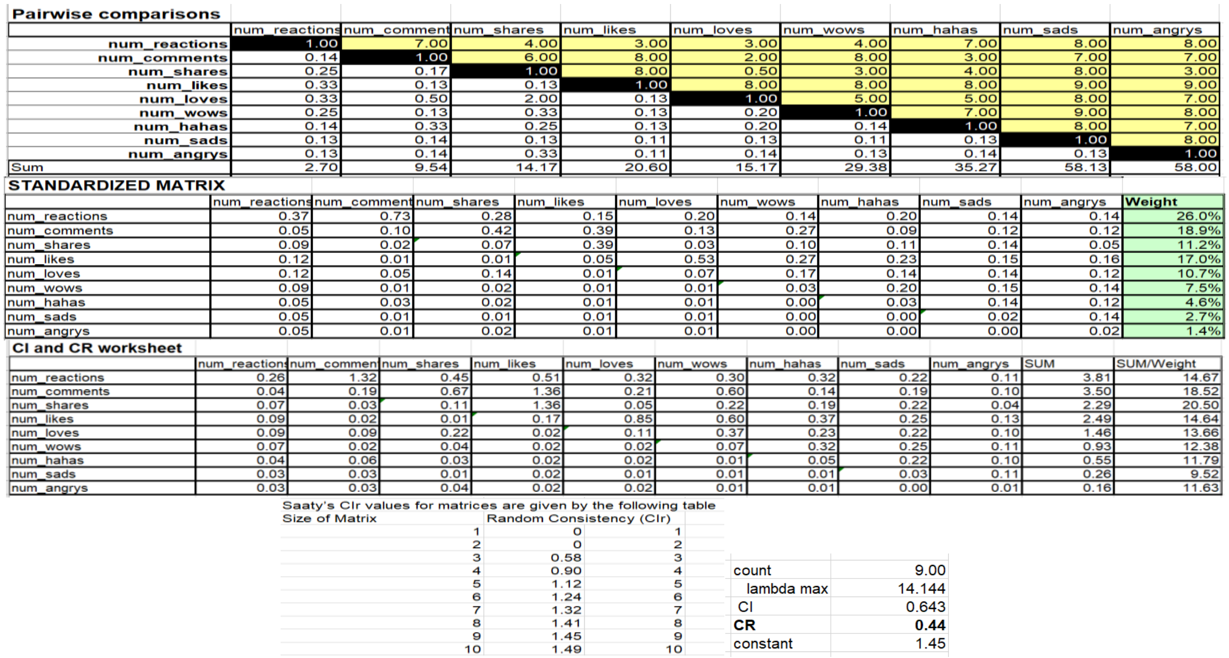
	statuses_count	followers_count	friends_count	favourites_count	listed_count	sex_code	lang_code	SUM	SUM/Weight
statuses_count	0.33	1.63	0.65	0.40	0.41	0.34	0.11	3.86	11.63
followers_count	0.05	0.23	1.14	0.92	0.16	0.30	0.16	2.96	12.76
friends_count	0.08	0.03	0.16	0.66	0.33	0.30	0.21	1.77	10.89
favourites_count	0.11	0.03	0.03	0.13	0.57	0.26	0.14	1.28	9.66
listed_count	0.07	0.12	0.04	0.02	0.08	0.26	0.16	0.74	9.14
sex_code	0.04	0.03	0.02	0.02	0.01	0.04	0.14	0.29	7.82
lang_code	0.07	0.03	0.02	0.02	0.01	0.01	0.02	0.18	7.89

Saaty's Clr values for matrices are given by the following table

Size of Matrix	Random Consistency (Clr)	
1	0	1
2	0	2
3	0.58	3
4	0.90	4
5	1.12	5
6	1.24	6
7	1.32	7
8	1.41	8
9	1.45	9
10	1.49	10

count	7.00
lambda max	9.968
CI	0.495
CR	**0.37**
constant	1.32

Fig. 19. Facebook Profile Credibility Weights (Model 1).



Pairwise comparisons

	num_reactions	num_comment	num_shares	num_likes	num_loves	num_wows	num_hahas	num_sads	num_angrys
num_reactions	1.00	7.00	4.00	3.00	3.00	4.00	7.00	8.00	8.00
num_comments	0.14	1.00	6.00	8.00	2.00	8.00	3.00	7.00	7.00
num_shares	0.25	0.17	1.00	8.00	0.50	3.00	4.00	8.00	3.00
num_likes	0.33	0.13	0.13	1.00	8.00	8.00	8.00	9.00	9.00
num_loves	0.33	0.50	2.00	0.13	1.00	5.00	5.00	8.00	7.00
num_wows	0.25	0.13	0.33	0.13	0.20	1.00	7.00	9.00	8.00
num_hahas	0.14	0.33	0.25	0.13	0.20	0.14	1.00	8.00	7.00
num_sads	0.13	0.14	0.13	0.11	0.13	0.11	0.13	1.00	8.00
num_angrys	0.13	0.14	0.33	0.11	0.14	0.13	0.14	0.13	1.00
Sum	2.70	9.54	14.17	20.60	15.17	29.38	35.27	58.13	58.00

STANDARDIZED MATRIX

	num_reactions	num_comment	num_shares	num_likes	num_loves	num_wows	num_hahas	num_sads	num_angrys	Weight
num_reactions	0.37	0.73	0.28	0.15	0.20	0.14	0.20	0.14	0.14	26.0%
num_comments	0.05	0.10	0.42	0.39	0.13	0.27	0.09	0.12	0.12	18.9%
num_shares	0.09	0.02	0.07	0.39	0.03	0.10	0.11	0.14	0.05	11.2%
num_likes	0.12	0.01	0.01	0.05	0.53	0.27	0.23	0.15	0.16	17.0%
num_loves	0.12	0.05	0.14	0.01	0.07	0.17	0.14	0.14	0.12	10.7%
num_wows	0.09	0.01	0.02	0.01	0.01	0.03	0.20	0.15	0.14	7.5%
num_hahas	0.05	0.03	0.02	0.01	0.01	0.00	0.03	0.14	0.12	4.6%
num_sads	0.05	0.01	0.01	0.01	0.01	0.00	0.00	0.02	0.14	2.7%
num_angrys	0.05	0.01	0.02	0.01	0.01	0.00	0.00	0.00	0.02	1.4%

CI and CR worksheet

	num_reaction	num_commen	num_shares	num_likes	num_loves	num_wows	num_hahas	num_sads	num_angrys	SUM	SUM/Weight
num_reactions	0.26	1.32	0.45	0.51	0.32	0.30	0.32	0.22	0.11	3.81	14.67
num_comments	0.04	0.19	0.67	1.36	0.21	0.60	0.14	0.19	0.10	3.50	18.52
num_shares	0.07	0.03	0.11	1.36	0.05	0.22	0.19	0.22	0.04	2.29	20.50
num_likes	0.09	0.02	0.01	0.17	0.85	0.60	0.37	0.25	0.13	2.49	14.64
num_loves	0.09	0.09	0.22	0.02	0.11	0.37	0.23	0.22	0.10	1.46	13.66
num_wows	0.07	0.02	0.04	0.02	0.02	0.07	0.32	0.25	0.11	0.93	12.38
num_hahas	0.04	0.06	0.03	0.02	0.02	0.01	0.05	0.22	0.10	0.55	11.79
num_sads	0.03	0.03	0.01	0.02	0.01	0.01	0.01	0.03	0.11	0.26	9.52
num_angrys	0.03	0.03	0.04	0.02	0.02	0.01	0.01	0.00	0.01	0.16	11.63

Saaty's Clr values for matrices are given by the following table

Size of Matrix	Random Consistency (Clr)	
1	0	1
2	0	2
3	0.58	3
4	0.90	4
5	1.12	5
6	1.24	6
7	1.32	7
8	1.41	8
9	1.45	9
10	1.49	10

count	9.00
lambda max	14.144
CI	0.643
CR	**0.44**
constant	1.45

Fig. 20. Facebook Status Message Credibility Weights (Model 2).

F. Data Modeling: Proposed Models

In this section, we will illustrate the classification Models performed to classify Users' Profiles into fake or genuine users and Credible or non-credible profiles, as seen in Fig. 21 and 22.

Model 1: Using Supervised Learning "Classification Model" (fake or genuine users).

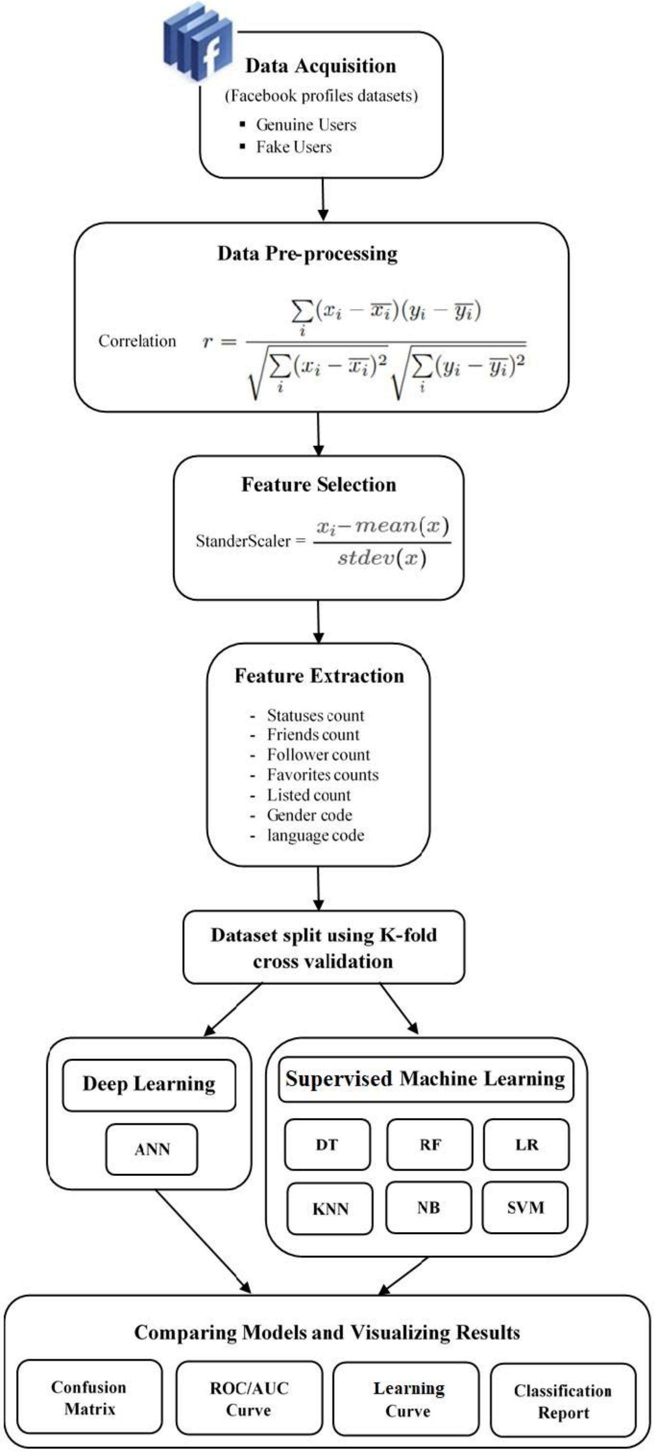

Data Acquisition
(Facebook profiles datasets)
Genuine Users
Fake Users
Data Pre-processing
Correlation $r = \frac{\sum_i (x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum_i (x_i - \overline{x_i})^2}\sqrt{\sum_i (y_i - \overline{y_i})^2}}$
Feature Selection
StanderScaler = $\frac{x_i - mean(x)}{stdev(x)}$
Feature Extraction
- Statuses count
- Friends count
- Follower count
- Favorites counts
- Listed count
- Gender code
- language code
Dataset split using K-fold cross validation
Deep Learning
ANN
Supervised Machine Learning
DT
RF
LR
KNN
NB
SVM
Comparing Models and Visualizing Results
Confusion Matrix
ROC/AUC Curve
Learning Curve
Classification Report

Fig. 21. Fake or Genuine Proposed Model.

Model 2: Using Unsupervised Learning "Clustering Model" (Credible or non-credible profiles).

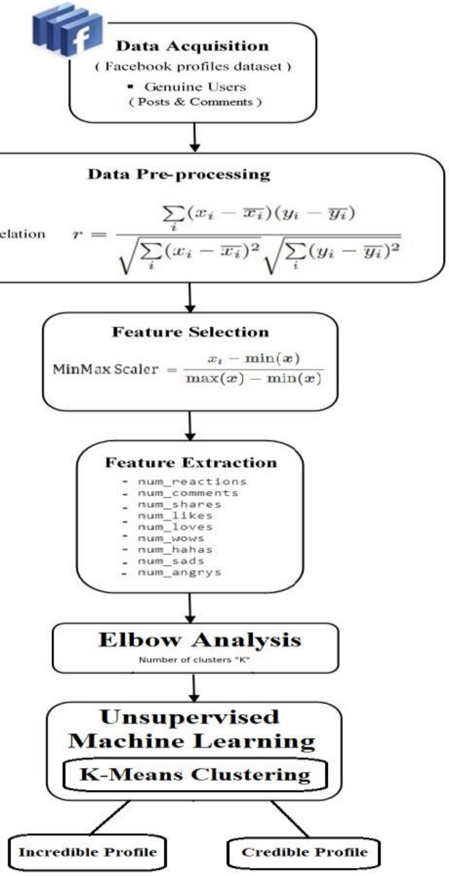
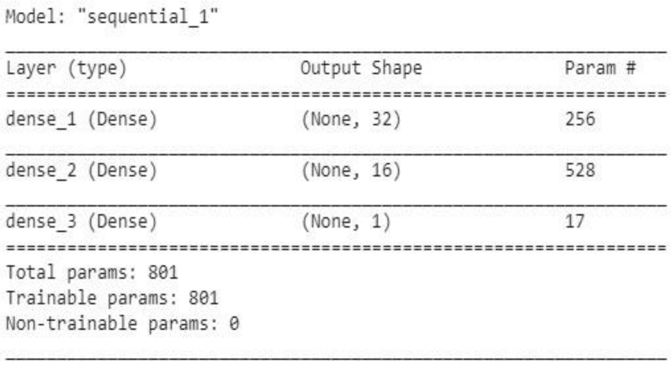


Fig. 22. Credible or Non-Credible Proposed Model.

G. ANN Model: Layers Summary

In our first model, we have built a deep neural network for binary classification to be able to model non-linear relationships and to use Feed-forward neural networks. The model implemented by using, 2 hidden layers with 32 and 16 nodes, using relu activation function. As seen in Fig. 23, the output layer employs the sigmoid activation since it is a binary classification problem. The model achieves 94% training accuracy, pretty well.



```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 32)                256
_________________________________________________________________
dense_2 (Dense)              (None, 16)                528
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 17
=================================================================
Total params: 801
Trainable params: 801
Non-trainable params: 0
_________________________________________________________________
```

Fig. 23. ANN Model: Layers Summary.

V. Experiment Results and Discussion

A. Experiment 1: Discussions

Algorithm 6: Pseudo-code for a Users' profile Binary Classification Model

```
Input: Datasets, Classifiers
Output: All Models Performance Analysis
Procedure:
Datasets → {Fake / Genuine, Non-Credible/Credible};
Classifiers → {RF, KNN, SVC, DT, LR, NB, NN};
AllAccuracyScores→{}:
AllRecallScores→{};
AllPrecisionScorcs→{};
Allf1-Scores→{};
AllAUCScores→{};
for DS ∊ Datasets do
    for Xtrain, Xtest ∊ (80%/20% split (DS)) do
        Xtrain, Xtest →Perform StanderScaler
        for clf ∊ Classifiers do
            clf →TrainClassifier(clf, XtrainLabels);
            predictions→(cls,Xtest);
            Accuracy→ ComputeAccuracy(predictions,XtestLabels);
            Recall→ ComputeRecall(predictions,XtestLabels);
            Precision→ComputePrecision(predictions,XtestLabels);
            F1-Score→ComputeF1-Score(predictions,XtestLabels);
            AUC→ComputeAUC(predictions,XtestLabels);
        end for
        return Learning Curves
        return Confusion Matrices (with/without normalization)
        return ROC/AUC Curves
        return Classification Reports
        (AllAccuracyScores, AllRecallScores, AllPrecisionScorcs,
         Allf1-Scores, AllAUCScores)
        return Fake / Genuine users' profile
        return Non-Credible / Credible users' profile
    end for
end for
```

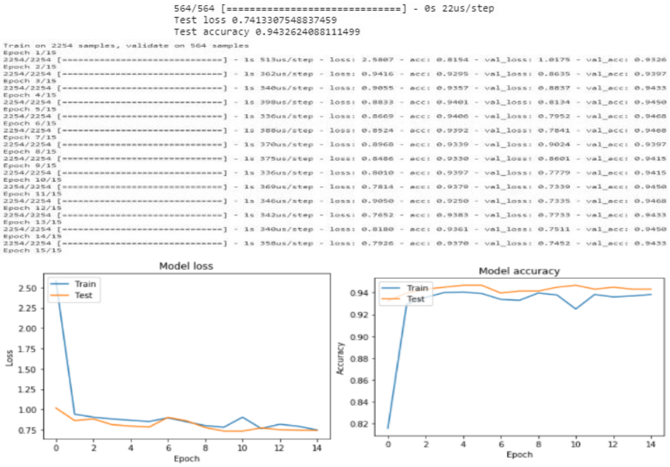
1) *Discussion on learning curves:* For model performance on training and testing, we plot Learning curves that graphs data against varying numbers of training instances. It allows training and testing performance to be viewed separately, to estimate how well models generalize to new data and allow diagnosis of bias and variance problems. High bias is when training/testing errors are high and converge, resulting in poor generalization. High variance is when there is a large gap between the errors, which could indicate there is not enough data or the model is too complex with too many features.

Fig. 24 and 25 represent the learning curve plots for each model in the first experiment. As illustrated in Fig. 24, a neural network learning curve showed a case of a good fit. The training loss plot decreased to a point of stability. Also, the validation loss plot decreased to a point of stability and as noticed from the curve, the gap between both is small.

2) *Discussion on confusion matrixes:* In the first experiment, the total number of observations that have been labeled was 564 observations in a size of 2x2 matrix according to the binary classification problem.

The following Fig. 26, 27, and 28 shows the confusion matrix for each classification classifier applied on the dataset.

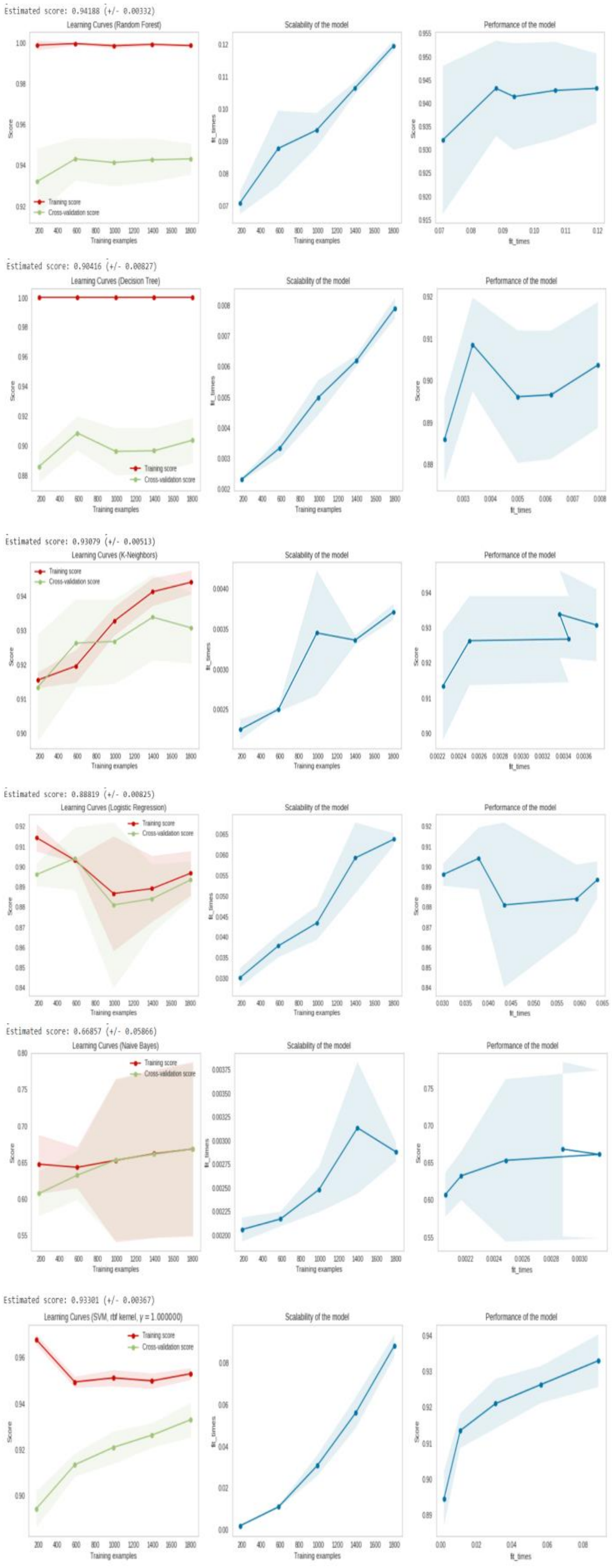
As shown in Fig. 27, the neural network confusion matrix for each of these four values has a specific name. The bottom right is called 'true positives' and indicates that 269 cases, predicted correctly by the classifier, showing the user with a genuine account. The upper left is called 'true negatives' and indicates that 263 cases the classifier correctly predicted the users with a fake account. The upper right is called 'false positives' and indicates that 5 cases only the classifier incorrectly predicted, and the user has a fake account, however, in fact, they do not. The bottom left is called 'false negatives' and indicates that in 27 cases the classifier incorrectly predicted that the user account is genuine when in fact they do have a fake account. We also use the confusion matrix to calculate the accuracy by adding the 'true positives' and the 'true negatives' then dividing them by the total number of observations.



```
564/564 [==============================] - 0s 22us/step
Test loss 0.7413307548837459
Test accuracy 0.9432624088111499
```

```
Train on 2254 samples, validate on 564 samples
Epoch 1/15
2254/2254 [==============================] - 1s 513us/step - loss: 2.5807 - acc: 0.8154 - val_loss: 1.0175 - val_acc: 0.9326
Epoch 2/15
2254/2254 [==============================] - 1s 362us/step - loss: 0.9416 - acc: 0.9295 - val_loss: 0.8635 - val_acc: 0.9397
Epoch 3/15
2254/2254 [==============================] - 1s 340us/step - loss: 0.9055 - acc: 0.9357 - val_loss: 0.8837 - val_acc: 0.9433
Epoch 4/15
2254/2254 [==============================] - 1s 398us/step - loss: 0.8833 - acc: 0.9401 - val_loss: 0.8134 - val_acc: 0.9450
Epoch 5/15
2254/2254 [==============================] - 1s 336us/step - loss: 0.8669 - acc: 0.9406 - val_loss: 0.7952 - val_acc: 0.9468
Epoch 6/15
2254/2254 [==============================] - 1s 388us/step - loss: 0.8524 - acc: 0.9392 - val_loss: 0.7841 - val_acc: 0.9468
Epoch 7/15
2254/2254 [==============================] - 1s 370us/step - loss: 0.8968 - acc: 0.9339 - val_loss: 0.9024 - val_acc: 0.9397
Epoch 8/15
2254/2254 [==============================] - 1s 375us/step - loss: 0.8486 - acc: 0.9330 - val_loss: 0.8601 - val_acc: 0.9415
Epoch 9/15
2254/2254 [==============================] - 1s 336us/step - loss: 0.8010 - acc: 0.9397 - val_loss: 0.7779 - val_acc: 0.9415
Epoch 10/15
2254/2254 [==============================] - 1s 369us/step - loss: 0.7814 - acc: 0.9379 - val_loss: 0.7339 - val_acc: 0.9450
Epoch 11/15
2254/2254 [==============================] - 1s 346us/step - loss: 0.9050 - acc: 0.9250 - val_loss: 0.7335 - val_acc: 0.9468
Epoch 12/15
2254/2254 [==============================] - 1s 342us/step - loss: 0.7652 - acc: 0.9383 - val_loss: 0.7733 - val_acc: 0.9433
Epoch 13/15
2254/2254 [==============================] - 1s 340us/step - loss: 0.8180 - acc: 0.9361 - val_loss: 0.7511 - val_acc: 0.9450
Epoch 14/15
2254/2254 [==============================] - 1s 358us/step - loss: 0.7926 - acc: 0.9370 - val_loss: 0.7452 - val_acc: 0.9433
Epoch 15/15
```

Fig. 24. Neural Network Learning Curve.



```
Estimated score: 0.94188 (+/- 0.00332)
Estimated score: 0.90416 (+/- 0.00827)
Estimated score: 0.93079 (+/- 0.00513)
Estimated score: 0.88819 (+/- 0.00825)
Estimated score: 0.66857 (+/- 0.05866)
Estimated score: 0.93301 (+/- 0.00367)
```

Fig. 25. Summarization for each Model Learning Curves.

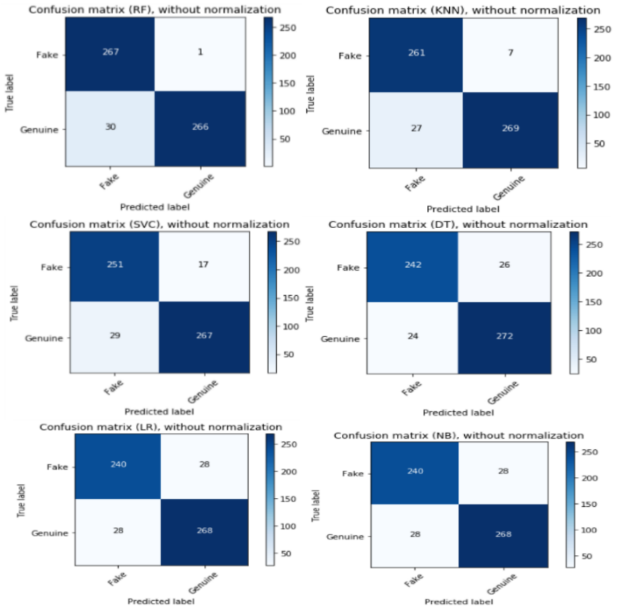
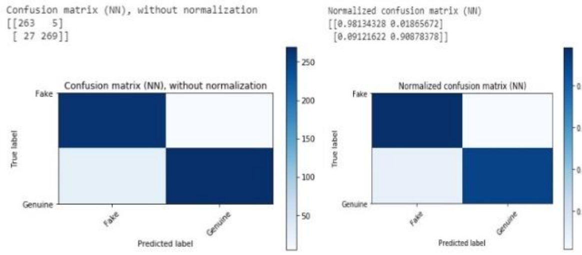


Fig. 26. Summarization for each Model Confusion Matrix (Model 1).



```
Confusion matrix (NN), without normalization
[[263   5]
 [ 27 269]]
```

```
Normalized confusion matrix (NN)
[[0.98134328 0.01865672]
 [0.09121622 0.90878378]]
```

Fig. 27. Neural Network Confusion Matrix (Model 1).

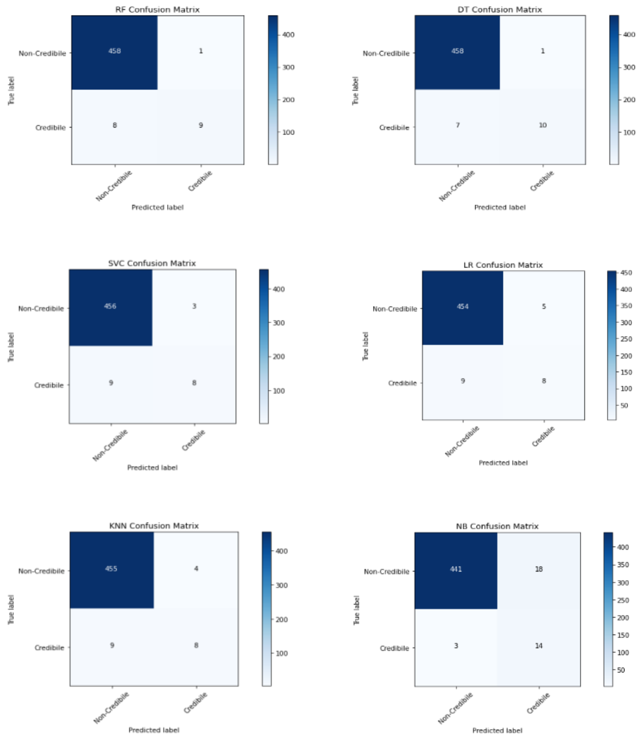


Fig. 28. Summarization for Credibility Confusion Matrix (Model 2).

3) Discussion on ROC/AUC Curves: AUC was calculated for each classifier and used to plot the ROC curve plots to compare the discriminatory powers of the models based on predicted outcome vs. true outcome, as illustrated in the below cumulative, Fig. 29. The ROC curve visualizes the ability to pick a threshold that balances both "sensitivity" and "specificity", to produce the model. Unfortunately, the thresholds cannot be viewed, that used to generate the ROC curve (on the curve itself.)

4) Discussion on models performance: As shown in Fig. 30, we have summarized all the accuracies, precisions, recalls, and f1-scores that had been achieved for each classification model in our binary classification study shown in (Experiment 1).

5) Discussion on credibility score: The best classifier with the best accuracy score in the classification report was the Random Forest classifier, which achieved 95% and the second-best accuracy score computed for the Neural Network model that achieved 94% in classifying users into the fake or genuine.

B. Experiment 2

In this experiment, the dataset had pre-processed using stemming, and stop-lists to vector the sets. We had performed the LDA (Topic Modelling algorithm) to generate the 10 topic vectors and assigned the most relevant topic label. With the generated 10 topic vectors, we had performed PCA to visualize the distribution, created a radar chart to visualize the distribution of sentiment emotion on each topic, and two correlational matrices to visualize the relationship between topics.

We also used the k-Mean clustering algorithm for user analysis and segmentation. Then, we analyzed and grouped users' profiles based on their number of behaviors like 'share' or 'comment' on posts, in addition to the number of sentiment reactions on those posts including 'like', 'love', 'wow', 'haha', 'sad', and 'angry'. This is useful to identify active and inactive users and classify profiles to credible and non-credible profiles, as seen in Fig. 31 and 32.

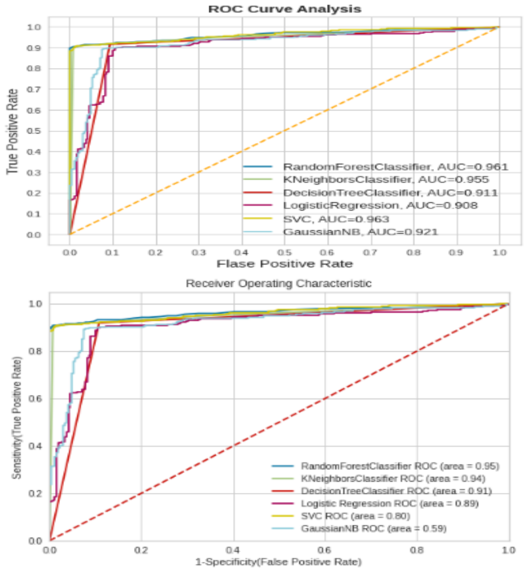
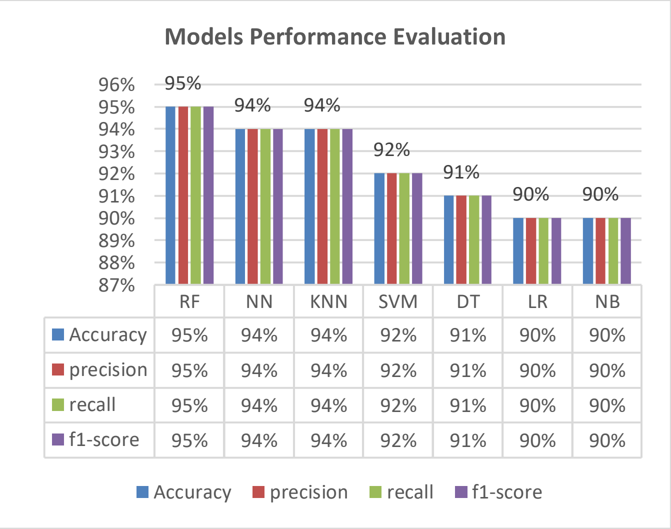
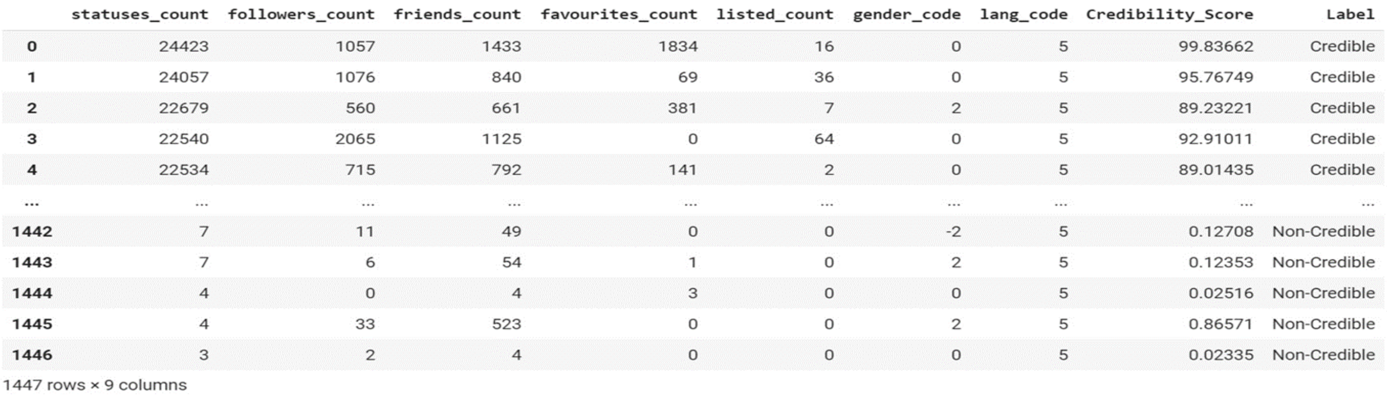


Fig. 29. Cumulative ROC / AUC Curve (Model 1& 2).



	RF	NN	KNN	SVM	DT	LR	NB
🟦 Accuracy	95%	94%	94%	92%	91%	90%	90%
🟥 precision	95%	94%	94%	92%	91%	90%	90%
🟩 recall	95%	94%	94%	92%	91%	90%	90%
🟪 f1-score	95%	94%	94%	92%	91%	90%	90%

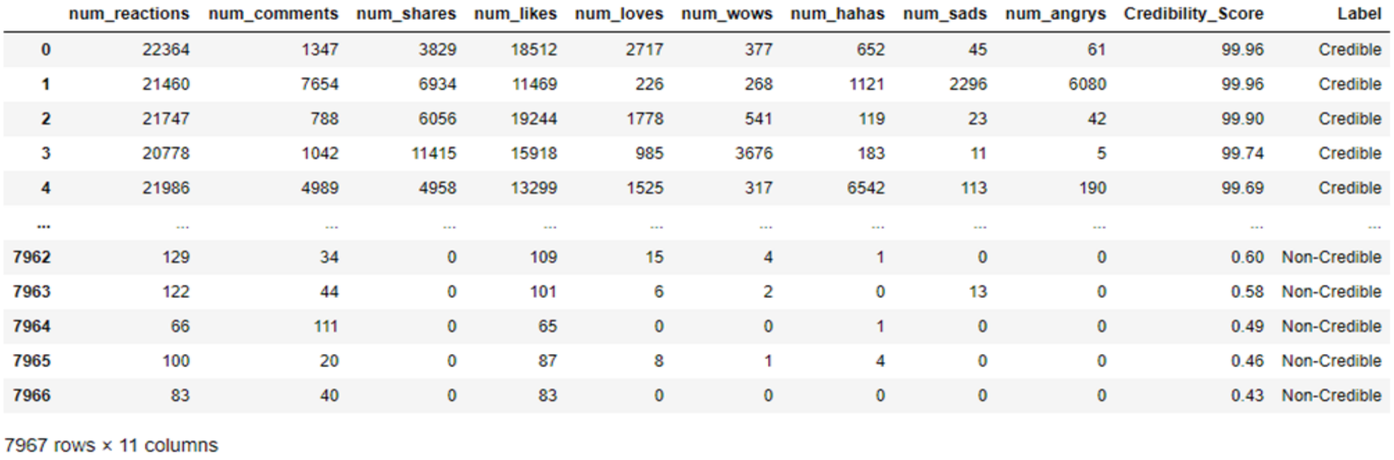
Fig. 30. Models Performance Evaluation (Model 1).



	statuses_count	followers_count	friends_count	favourites_count	listed_count	gender_code	lang_code	Credibility_Score	Label
0	24423	1057	1433	1834	16	0	5	99.83662	Credible
1	24057	1076	840	69	36	0	5	95.76749	Credible
2	22679	560	661	381	7	2	5	89.23221	Credible
3	22540	2065	1125	0	64	0	5	92.91011	Credible
4	22534	715	792	141	2	0	5	89.01435	Credible
...	...	...	...	...	...	...	...	...	...
1442	7	11	49	0	0	-2	5	0.12708	Non-Credible
1443	7	6	54	1	0	2	5	0.12353	Non-Credible
1444	4	0	4	3	0	0	5	0.02516	Non-Credible
1445	4	33	523	0	0	2	5	0.86571	Non-Credible
1446	3	2	4	0	0	0	5	0.02335	Non-Credible

1447 rows × 9 columns

Fig. 31. Exploratory Samples of users' Profile Credibility Score.



	num_reactions	num_comments	num_shares	num_likes	num_loves	num_wows	num_hahas	num_sads	num_angrys	Credibility_Score	Label
0	22364	1347	3829	18512	2717	377	652	45	61	99.96	Credible
1	21460	7654	6934	11469	226	268	1121	2296	6080	99.96	Credible
2	21747	788	6056	19244	1778	541	119	23	42	99.90	Credible
3	20778	1042	11415	15918	985	3676	183	11	5	99.74	Credible
4	21986	4989	4958	13299	1525	317	6542	113	190	99.69	Credible
...	...	...	...	...	...	...	...	...	...	...	...
7962	129	34	0	109	15	4	1	0	0	0.60	Non-Credible
7963	122	44	0	101	6	2	0	13	0	0.58	Non-Credible
7964	66	111	0	65	0	0	1	0	0	0.49	Non-Credible
7965	100	20	0	87	8	1	4	0	0	0.46	Non-Credible
7966	83	40	0	83	0	0	0	0	0	0.43	Non-Credible

7967 rows × 11 columns

Fig. 32. Exploratory Samples of user Reaction Credibility Score.

VI. Conclusion

In this paper, we have implemented two experiments on Facebook user profiles with content generated as posts or comments on pages as CNN page. The first experiment is a binary classification model that automatically detects the fake and genuine profiles. Then, real users classified to credible or not each according to credibility score computed. The researchers computed the credibility score based on the credibility theory using the AHP approach to compute the weights of the correlated features. In this experiment, the Machine Learning and Deep Learning pipeline had been followed. Utilized six supervised machine learning classifiers such as SVM, RF, DT, KNN, LR, NB, and a Deep Learning NN model. The second experiment is a clustering model that classifies the users into two groups of clusters to identify the credible and non-credible users according to their behaviors on posts and comments. In the second model, we had extracted 10 sets of topics by using LDA. After that, we visualized them with sentiments emotions counted from the status message using a correlation matrix to show the dependence and relationship between these various sets. We used the radar charts to plot the 10 sets of topics with sentiment emotions features. Then, we found that the most reactions related to sadness or angry as a negative behavior response related to the time the dataset collected concerning political directions and presidential elections. We had verified the results of the observations for each emotional reaction, then visualized and computed the Principal Component Analysis. In this experiment, we also followed the Machine Learning pipeline using the k-means cluster as an unsupervised learning algorithm to assign each status to the most relevant topic creating the topics sets. And we used the supervised learning algorithms to classify the labels for the topic’s sets. In addition to experiment 1, we have plotted the Learning Curves for each model performed to show the model stability. Applied different methods to evaluate the performance of the classifiers, such as the Confusion Matrix table and the ROC/AUC curve. Those two methods described the classifier performance. Implementing both whenever possible, will be beneficial for evaluating any model. The primary characteristic of the Confusion Matrix is the numerous evaluation that can be calculated with it, such as Accuracy Score, Precision Score, Recall Score, and the F1-Score. Also, we can concentrate on the metrics that resemble our research scope. On the other hand, the major characteristic of ROC/AUC curves is, they do not demand us to pick a classification threshold, unlike the Confusion Matrix. We also notice that the main difference between machine learning and deep learning is that deep learning merge’s the feature extraction with classification in one process and we don’t need to apply the full analysis phase. At the end of this study, experiment ‘1’ results achieved 95% by using the RF classifier and achieved 94% by using the NN model to classify fake and genuine users. Experiment ‘2’ classified the user profiles into credible and non-credible users. This work considered to be the first step that should be performed to measure the profile credibility on Social Media “Facebook” especially status messages with sentiment emotions responses.

References

[1] Gupta, A., & Kaushal, R. (2017). Towards detecting fake user accounts in Facebook. In 2017 ISEA Asia Security and Privacy (ISEASP) (pp. 1-6). IEEE.

[2] Saikaew, K. R., & Noyunsan, C. (2015). Features for measuring credibility on Facebook information. International Scholarly and Scientific Research & Innovation, 9(1), 174-177.

[3] Khedr, Ayman E., N Yaseen, (2017), stock market behavior using data mining technique and news sentiment analysis, International Journal of Intelligent Systems and Applications 9 (7), 22-30.

[4] Khedr, A. E., Idrees, A. M., E Shaaban, E., (2020). Automated Ham-Spam Lexicon Generation Based on Semantic Relations Extraction, International Journal of e-Collaboration (IJeC), 16 (2), 45-64.

[5] Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. " O'Reilly Media, Inc.".

[6] Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media.

[7] Harrington, P. (2012). Machine learning in action. Manning Publications Co.

[8] Sultan, Torky I., Khedr, Ayman E., Nasr, Mona M., and Ismail, Walaa S., (2012). Semantic Interoperability-Traditional and Ontology-Based Approaches. In the 1st International Conference of Computing and Informatics (ICCI'2012).

[9] Sharaf Eldin, Ahmed, Khedr, Ayman E., Al-Sharif, Fahad Kamal, (2015), Cross-Language Semantic Web Service Discovery to Improve

the Selection Mechanism by using Data Mining Techniques, International Journal of Computer Applications, 110 (2), 0975 – 8887.

[10] Lê, N. C., Dao, M. T., Nguyen, H. L., Nguyen, T. N., & Vu, H. (2019). An Application of Random Walk on Fake Account Detection Problem: A Hybrid Approach. arXiv preprint arXiv:1911.07609.

[11] Smruthi, M., & Harini, N. (2019). A Hybrid Scheme for Detecting Fake Accounts in Facebook. International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-7, Issue-5S3.

[12] Wani, S. Y., Kirmani, M. M., & Ansarulla, S. I. (2016). Prediction of Fake Profiles on Facebook using Supervised Machine Learning Techniques-A Theoretical Model. International Journal of Computer Science and Information Technologies (IJCSIT), 7(4), 1735-1738.

[13] Yao H., Jiang C., Qian Y. (2019) Introduction. In: Developing Networks using Artificial Intelligence. Wireless Networks. Springer, Cham.

[14] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160, 3-24.

[15] Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems, 83-124.

[16] Tan, P. N., Steinbach, M., & Kumar, V. (2006). Classification: basic concepts, decision trees, and model evaluation. Introduction to data mining, 1, 145-205.

[17] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[18] Fosseng,, S. (2013). Learning Distance Functions in k-Nearest Neighbors (Master's thesis, Institutt for datateknikk og informasjonsvitenskap).

[19] Steinbach, M., & Tan, P. N. (2009). kNN: k-nearest neighbors. The top ten algorithms in data mining, 151-162.

[20] Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.

[21] Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.

[22] Kampakis, S. (2018). What deep learning is and isn’t.

[23] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the royal statistical society. series c (applied statistics), 28 (1), 100-108.

[24] Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition, 30 (7), 1145-1159.

[25] McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. The American Statistician, 32 (1), 12-16.

[26] Pearson, K. (1901). Principal Components Analysis. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 6 (2), 559.

Agent Mining Framework for Analyzing Moroccan Olive Oil Datasets

Belabed Imane[1], Jaara El Miloud[2], Belabed Abdelmajid[3], Talibi Alaoui Mohammed[4]
University Mohammed the First, Oujda, Morocco[1, 2, 3]
Faculty of Science and Technology, Fez, Morocco[4]

***Abstract*—Data mining and intelligent agents have become two promising research areas. Each intelligent agent functions independently while cooperating with other agents, to perform effectively assigned tasks. The main goal of this research, is to provide a mining implementation that can help biological researchers for discovering parameters that affect the cost of olive oil in Morocco. To solve this problem, we used a method involving two data mining techniques, clustering of variables, quantitative association rules and multi-agent system to fuse these two techniques. Therefore, we have developed a multi-agent framework that has been validated by using concrete data from the Provincial Direction of Agriculture of Berkane, Morocco. To prove the performance of our framework, we tested the proposed multi-agent tool using three datasets from different fields. Conforming to biological researchers, our method generates a clear knowledge because the framework proposes high-confidence rules that can correctly identify olive oil factors.**

Keywords—Quantitative association rules; clustering of variables; multi-agent system

I. Introduction

Data mining technology aims to find useful knowledge from the database. In fact, data mining technology engages a crucial role in many business analysis and prediction applications used to complete data analysis.

The clustering covers many various algorithms and methods for grouping similar kinds of objects into various categories. Such algorithms or methods are associated with organizing the observed data into expressive structures. If two objects belong to the same group the correlation between them is the largest, otherwise it is the smallest. In light of the above, cluster analysis can be used to discover structures in data without providing explanations. This research takes the partition based clustering such K-Means algorithm for variables.

The issue of finding relevant relationships between attributes has considerably studied through the association rules concept. Association rules are a technique that allows users to discover associations between different objects in databases. The results are given in the form of antecedent and consequent. The reliability of rules is usually measured by using statistical function, as for instance the support and the confidence. The rules that maximize the support and the confidence are more considered in the mining tasks.

The execution of association rules on numeric attributes, is problematic, as a conventional approach, the most methods perform a pre-processing phase called discretization, which refers to the process through which we can transform continuous variables, into a discrete form before executing the learning task.

To address this issue, in this work, we recommend the use of genetic algorithm to process quantitative datasets. Consequently exceed the discretization phase. The adoption of a genetic algorithm is sustained by our previous work, in which we evaluate the association rules obtained by running apriori and genetic algorithm using quantitative datasets in multi-agent environment. The experiments show that genetic algorithm avoids the redundant rules with satisfied execution time [22].

The most motivation of this research is to obtain an automatic solution that consolidates the two data mining techniques illustrated before. Therefore, the idea of integrating multi-agent technology into data mining applications turns to be useful. In fact, it seeks supplementary benefits by integrating agent technology into data mining technology. This interaction allows us to create an effective solution to solve the main problems of this work, which includes the discovery of factors that increase the profit of Moroccan olive oil and thus reduce the cost of olive oil.

The rest of the paper is split as follows. Section II discusses the data mining process adopted to implement the agent framework, while Section III briefly presents the data mining techniques used. Section IV presents an overview of multi-agent system concept. Section V presents the works related to our research area. Section VI details the architecture of the agent-mining tool, and presents the implementation process. Section VII illustrates the results obtained by applying the developed tool on three test cases specified in Table I. Section VIII outlines the use of the framework in the Moroccan agricultural domain test case. Section IX ends the paper and highlights suggestions for future work.

A. *Motivation and Our Contribution*

Over the last years, Morocco has developed a new agricultural strategy called Green Morocco Plan. The project was aimed at supporting modern agriculture, with benefit and high productivity that correspond to market requirements. The plan encourages private investments, in order to improve the chain of productivity and develop industrial activities related to agriculture.

The Green Morocco Plan is an agricultural strategy launched in 2008 that aims to make agriculture, the main growth engine of the national economy over the next ten to fifteen years, with significant benefits in terms of growth. New instruments were designed for the Green Moroccan Plan implementation. The plan is structured around seven pillars. Most actions come under Pillar I and Pillar II, the goal of Pillar I of the Green Morocco Plan is the development of agriculture with high benefit and high productivity. This requires the voluntarism creation of agricultural development poles insuring high benefit and market requirements [2].

The pillar I aims at improving the production chain of high benefit crops such olive [1].

The goal of this paper is to deal with Pillar I, particularly the identification of the factors that optimize the Moroccan olive oil cost. Our research consists on creating an efficient process to discover theses parameters and support decision making in this field.

II. Related Work

In this section, we will briefly introduce previous researches in machine learning area including knowledge discovery in agent systems or multi-agent systems. The techniques used in these studies include association rule mining, clustering mining, and rule generation algorithms. The proposed approach is mainly related to two areas of research, knowledge extraction from dataset and knowledge modelling using the multi intelligent agent system. Popa proposed an intelligent recommendation system based on multi-agent, called Agent Discover. The purpose of the system is to solve the complexity of knowledge discovery and provide a tool to support researchers and non-expert users to explore knowledge discovery methods and quickly find results in the field [14]. Tong, proposed a real time Data Mining and Multi-Agent System called DMMAS. The DMMAS method uses data partitioning and multiple agents, and can choose to use heterogeneous or homogeneous data mining technology. Agent-based distributed processing can model and combine the results of all agents to improve the efficiency [15]. Nahar presented a paper in which she executes association rule mining and a computational intelligence to recognize the factors, which contributes to the apparition of heart diseases for males and females. This study proposed an experiment using three rule generation algorithms Apriori, Predictive Apriori and Tertius to extract rules from heart disease data [16].

Ait-Mlouk proposed a method based on multi-criteria analysis to discover a category of relevant association rules. The author uses multi-agent system to integrate, manage, and model the quality measurement according to six agents working in cooperation [17]. Kaur produced a spatial data mining techniques to extract implicit knowledge from spatial attributes. These techniques are applied in different fields such as healthcare, marketing, and remote sensing databases to improve planning and decision-making process [18].

Salleb-Aouissi, proposes a system based on generating association rules from quantitative datasets by using the concept of genetic algorithm. They tested their tool on both real datasets from medical domain and synthetic datasets [21].

The approach presented in this work, consists on providing an agent mining tool that includes two data mining techniques represented by two categories of mining agents. Thanks to the integration of genetic algorithm, the present solution uses one agent in the association rules phase. Thus, processes efficiently the issue of redundant rules presented in Ait-Mlouk [17]. In fact, in that work the authors use a multi criteria approach to filter significant rules. Consequently, they use six agents to generate association rules.

In addition, our proposed agent tool overcomes the Salleb-Aouissi approach mainly in the execution time. Due to the integration of both multi-agent system and K-Means clustering, the global execution time of every experiment is fourteen times lower than the best execution time in that work, for real datasets [21].

III. Data Mining Analysis Process

When we studied the problem of extracting knowledge from the Moroccan olive dataset, we request a solution based on the combination of clustering and association rules [3]. We recommend applying rule association algorithms to each of the defined clusters.

To achieve this goal, we choose K-Means for the clustering phase and genetic algorithm for association rules phase.

Among the knowledge discovery process, we confront two challenges such:

- Executing K-Means algorithm for variables.
- Extracting association rules from quantitative data.

To exceed the first constraint, we choose to process only quantitative datasets and transpose the data in order to cluster variables instead of observations. Then, we integrate K-Means algorithm using the Weka open source framework [4]. For the association rules phase, we chose to use genetic algorithm for quantitative association rules. The algorithm relies on genetic algorithm to find dynamically the optimal interval for numerical data.

The data used in this research are divided into two parts. The first one consists of using datasets from internet, uci-archives for machine learning [20], in order to validate the framework and evaluate the performance. The second part consists of analysing olive oil datasets, based on our framework. Table I, presents the datasets used in our experiments.

TABLE I. Data Specification

Data field	Nature of variable	Volume of variables
Absenteeism data	Quantitative	19 variables
Bio-degradation data	Quantitative	41 variables
Air quality data	Quantitative	12 variables
Olive crops datasets	Quantitative	17 variables

IV. Data Mining Algorithms

In this section, we highlight the data mining techniques used.

A. K-Means Clustering

The K-Means algorithm based on division is a popular clustering algorithm. This unsupervised algorithm is used commonly, for data mining and pattern identification. The algorithm aims at grouping data of more or less similarity by finding iteratively the centroid between the elements. In this work, the distance metric adopted is the Euclidian distance. The algorithm works as follow:

Step 1: Select a number of clusters, k.

Step 2: Choose k, the initial starting values to be the initial centroids.

Step 3: Affect each point to the cluster whose centroid is nearest to it in term of Euclidian distance.

Step 4: When each point is assigned to a cluster, recalculate the new k centroids.

Step 5: Repeat steps 3 and 4 until no point changes its cluster assignment, or until a maximum number of iterations is reached.

B. Association Rules

Association rule is a frequent data mining technique usually used for discovering relevant relationships between variables in large databases [5]. It studies the frequency of items that appear together in the transaction database and identifies frequent item sets based on the first threshold called support. The second threshold is called confidence, which calculates the conditional probability that an item shows up in a transaction when another item appears. The form of the association rules is: A B [s, c], where A and B are conjunctions of attribute value-pairs, and s, for support, is the probability that A and B show up together in a transaction and c, for confidence, is the conditional probability that B appears in a transaction when A exists.

C. Genetic Algorithm Overview

Genetic algorithm GA is an adaptive method that can be used to solve search and optimization problems, which is based on biological genetic processes.

The algorithm is based on the genetic processes of biological organisms. Over many generations, natural populations evolve confirming to the principles of natural selection and survival of the fittest first. By imitating this process, genetic algorithms can evolve solutions to real-world problems. The genetic algorithm uses multiples solutions collectively known as population. These solutions are usually coded in binary strings. Every solution or individual is assigned a fitness, which matches with the objective function of the search. Thereafter, the individual populations were modified to new populations by applying three operators like natural genetic operators, namely reproduction, crossover, and mutation.

Nevertheless, genetic algorithms have at common the following elements: population, selection according to fitness, crossover to reproduce new offspring, and random mutation of new offspring.

Initial population: The generation of initial population is executed as follow: Firstly, we consider the interval $[a_i, b_i]$, which represents the domain of quantitative variable a^{th}. Then the length of interval is decreasing until attaining a minimum support specified by the user. We note that the bounds a_i and b_i are chosen at random. This enables starting with enough diversity within the initial population.

The mutation operator sustains the diversity within the population. Indeed, it uses a selected rate, which determines the degree of changes to be made. The change is performed in term of modifying the length of the interval of selected variable [6].

The crossover operator is equivalent to reproduction in biological crossover. It is founded on taking two individuals, called parents [6], and generating new individuals. In our case, the concept is applicable to variable. Thereby, the interval of each variable is either inherited from one parent or formed by merging the bounds of the two parents [7].

Fitness function: A fitness function is a type of objective function that determines the optimally of chromosome in traditional genetic algorithm [8]. In this paper, we filter rules by measuring the support and the confidence. Therefore, the fitness function is built by combining the support and confidence parameters. We predefine the thresholds of minimum support, min_sup, and the minimum confidence, min_conf, for the algorithm [9].

D. Algorithm

The algorithm selects high confidence rules. In fact, the algorithm starts with a set of predefined rules in term of the implication direction. In this work, both the condition and the conclusion parts of a rule are recognized. In fact, the variables that constitute the conditions are defined in the clustering phase. Additionally, for every dataset, we choose only one variable to represent the conclusion part. The characteristics of these variables are illustrated in Table II.

The algorithm finds the optimal interval for quantitative variables present in that rule template. An optimization criterion is applied to select only the rules that maximize the support and the confidence parameters.

TABLE II. Variables to Predict

Data field	Variable to predict
Absenteeism data	Absenteeism hours
Bio-degradation data	Number of Halogen atoms
Air quality data	Air humidity
Olive crops data	Olive oil cost.

The algorithm follows the prototype of traditional genetic algorithm.

The algorithm inputs are the minimum confidence (min_conf), the minimum support (min_sup), the population size (Popsize), the crossover rate (Cross), the number of generations (Genum) and the mutation rates (Mut).

Genetic algorithm pseudo-code:

```
Select a set of attributes
Let $Rt$ a set of predefined rule specified on selected variables
For each $r \in Rt$ do
Choose the initial population P of Popsize
While $i \leq Genum$ do
Breed new generation through mutation and crossover operators i.e. Mut and Cross.
Extract the itemsets that deliver the best fitness to form the association rule values
i++
Return R= max (fitness (r)); r belongs to P.
```

V. Agent and Multi-Agent Systems

Various definitions have been proposed for the concept of multi-agent system (MAS). A multi-agent system is an approximately coupled network of problem-solver entities that collaborates to answer the issues related to the individual capabilities or knowledge of each entity.

Although the agents in a multi-agent system are often equipped with behaviors designed in advance, they often need to apprentice new behaviors online in order to improve progressively the performance of agents; consequently, the entire multi-agent system is ameliorated.

One of the current sectors of popularity of multi-agent systems is the smart agricultural domain especially the concept of internet of things where agents interact with each other to achieve their individual or shared goals. To perform in such an interactive environment, agents must surpass two challenges: they have to be ready to localize each other, since agents might appear, disappear, or change their status at any time. Additionally, the agents must be able to interact with each other [10].

A. Agents

An agent is defined as a computer system located in some environment and capable of acting autonomously in this environment to achieve its design goals. They can be reactive to changes produced in their environment; also, they are able to communicate and use computational intelligence to reach their goals by being proactive [11].

B. Multi-Agent Systems

By associating varied agents in one system to solve a problem, the produced system is called multi-agent system. These systems hold multiple agents that individually solve problems. They can communicate with each other and assist each other in achieving larger and more complex goals [12]. Multi-agent systems have been used in predicting the stock market, industrial automation etc. In this work, we developed a multi-agent system for predicting the factors that optimize the olive oil cost.

C. Multi-Agent Systems and Data Mining

Data mining and multi-agent systems present attractive features to form systems that are more intelligent [13].

- The combination of multi-agent autonomy and knowledge data mining provides adaptable systems.
- Data mining techniques such as association rule extraction have no equivalent in agent systems. Currently, these techniques deliver agents with the ability of learning and discovering.
- Data mining can enhance the agent capability of handling uncertainty via historical event analysis, dynamic mining, and active learning. By mining agent behavioural data, it is possible to reach a balance between agent autonomy and adequate learning.

VI. Multi-Agent System for Data Mining System

A. Architecture

As described below, the proposed mining framework shown in Fig. 1 includes four types of agents:

1) User agent.
2) Data agent.
3) Coordinator agent.
4) Clustering agent.
5) Association rules agent.

The agents are distributed in set of various containers taken from JADE, Java Agent Development Framework, in which our framework is implemented. One of these containers is the main container, which holds an AMS, Agent Management System and a DF, directory facilitator. The AMS agent is employed to manage the life cycle of other agents in the platform, and therefore the DF agent provides agent search services, such as yellow pages.

Fig. 1 presents a view of the framework architecture implemented in JADE. It shows the various categories of agents and their interactions along with the data mining algorithms. The figure illustrates the main container that holds the coordinator agent, additionally to the AMS agent and the DF agent as illustrated before. The other containers hold the specialized agent, the clustering agent, the data agent, the user agent and the association rules agent.

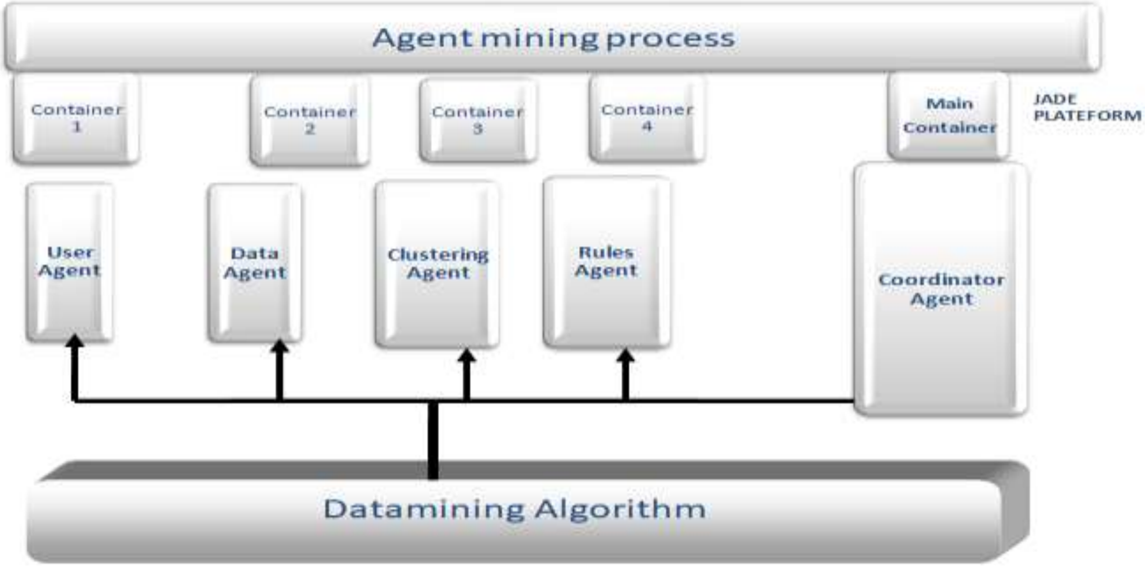


Fig. 1. Framework Architecture.

As shown in Fig. 1, the agent mining process is composed of multiple collaborative agents that act in several levels.

- **User agent** constitutes the interface between end user and the mining framework. The agent is responsible for obtaining the K value of K-Means and the datasets path.
- **Data agent** is responsible for charging the datasets from the data source and keeping meta-data information about the data source. There is a direct liaison between a data agent and a selected data source. Data agent is in charge of forwarding their data, when requested.
- **Coordinator agent** ensures the proper transmission of messages among the agents. It collects the user specifications and sends them to the corresponding agent.
- **Clustering agent** executes K-Means algorithm. Once the clustering agent have complete it task, it informs the coordinator agent.
- **Association rules agent** is responsible for generating supervised rules inside each predefined cluster through the genetic algorithm.

VII. Multi-Agent Framework Implementation

The present work proposes the implementation of the multi-agent system for data mining framework including clustering and association rules. We have developed java platform through Agent-Oriented Programming paradigm, AOP. The communication inter-agent is maintained through the recommendations of the standard Foundation for Intelligent, Physical Agents, FIPA. We have implemented our proposed tool with JADE [19].

JADE is FIPA-compliant middleware that enables the development of distributed applications based on the agent paradigm and is adequate to process large amounts of data with a data mining approach. JADE allows portability, which is assured by the use of Java, and defines an agent platform comprising a set of containers that may be distributed across a network. In the JADE platform, the main container holds a number of mandatory agent services, such as the AMS and DF agents. The DF agent is responsible for the yellow pages service at the main container of the JADE framework. In addition, the AMS agent is used to control the life cycles of other agents in the platform.

JADE holds two main products, a FIPA-compliant agent platform and a package to develop Java agents. JADE also provides the implementation of FIPA Agent Communication Language, ACL, which is a message-based protocol defined by FIPA.

Because this implementation requires a data-mining algorithm, the first implementation measure is to integrate the data-mining algorithm into the JADE platform. As a data mining tool, we choose the Weka 3.6.1 Java Library in the first phase of K-Means clustering.

VIII. Experimental Results

In this section, we expose the ability of our framework to discover knowledge with test cases in different domains, particularly in environment and agriculture. We will focus on the advantage provided by multi-agent technology to promote the data mining performance. Thus, for every phase of our data mining process we will present the results obtained by executing the three datasets illustrated before in Table I. Concretely, the performance evaluation during the data mining process, is measured through the following metrics:

- *Memory used:* physical memory consumed by the task of the agent when it has been executed. The resulting value is given in megabytes (MB).
- *The execution time with agent:* The amount of time the task took to process by the agent. The resulting value is given in seconds (s).

A. Clustering Phase

Clustering is conducted with an experimental K value of K-Means set into three. Then the framework was executed with two modes: with agent and without it, in order to evaluate the impact of multi-agent implementation. The number of clusters and the number of association rules agents vary proportionally. In fact, if the number of clusters is significant, more than one association rules agent is needed. However, the proposed architecture is extensible to include high number of clusters if the number of association rules agents is increased. In that case, several agents should be added such, a new manager agent that should firstly, orchestrates the interaction inside the association rules agents groups and secondly manage the communication with the coordinator agent.

Table III presents the result of applying the multi-agent concept and it consequent performance in the clustering phase. Due to the integration of multi-agent concept, the execution time is reduced greatly. However, in term of memory consumption, the agent platform consumes more than the java environment. The increase of the memory consumption is due to the DF agent hold in the main container. In fact, the DF agent should store it catalogue in memory. It contains, the yellow pages, which include the matches between the description of the agents and their proposed services.

B. Association Rules Phase

As discussed in Section I. The process of data mining used in this work, focuses on extracting rules between each discovered cluster and the variables to predict illustrated in Table II. This concept is valid for the three data source, bio-degradation data, absenteeism data, and air quality data.

The experiments are conducted with the genetic algorithm using quantitative datasets. The experimental thresholds of genetic algorithm are set to be as follows:

- The minimum confidence threshold fixed to 60%.
- The minimum support threshold fixed to 10%.
- The population size fixed to 250.
- The crossover rate fixed to 50%.
- The mutation rate fixed to 40%.

We chose to start the genetic algorithm with relatively low thresholds in order to generate more rules. Consequently, we can evaluate correctly the robustness of our tool.

The results presented below, concern the performance of the multi-agent approach integration, for rules mining phase using genetic algorithm. Concretely, the next subsections illustrate the process of rules extractions with two different modes, with agent and without. Note that, the number of rules extracted depends only on the genetic algorithm and not on the integration of the multi-agent technology.

1) Biodegradation datasets: Table IV presents the results of the association rules phase. For bio-degradation datasets, the rules extracted with genetic algorithm in each cluster were respectively 80 rules in cluster 1, 79 rules in cluster 2 and one rule in cluster 3. This experiment presents a few runtime improvements with the agent integration approach. However, the multi-agent platform is memory intensive compared to the java environment.

2) Air quality datasets: Table V presents, the three cluster results, with their execution time and memory consumption. The performance runtime is increased slightly in cluster 1 and cluster 3. However, for the air quality datasets, the multi-agent platform is more memory consuming than the traditional approach. The rules extracted with genetic algorithm in each cluster were respectively 42 rules in cluster 1, 15 rules in cluster 2 and 42 rule in cluster 3.

3) Absenteeism datasets: Table VI illustrates, the generation of rules using genetic algorithm for the absenteeism data. The multi-agent platform presents consistent improvement of the execution time, especially in cluster 1. In fact, the execution time is fifteen times less than the traditional approach. Nevertheless, the multi-agent platform stills memory intensive.

In conclusion, of this part, Tables IV, V and VI present the three experiment tests included execution time, memory consumption and number of generated rules.

In the association rules phase, the results obtained by integrating the agent approach on the three datasets, show an insignificant improvement in the execution time, except for the absenteeism test case, cluster 1, where the multi-agent approach presents an interesting runtime improvements. We note that, in this work the number of clusters is approximately small. The benefit of the multi-agent approach will be significant in the case of large number of clusters. In that case, the sum of the execution times obtained in each cluster would be considerable. Nevertheless, in this phase the increase of the memory consumption is always present.

TABLE III. RESULT OF CLUSTERING PHASE

K=3	**Data mining process clustering phase**		
Data	**Volume**	**Clustering with agent**	**Clustering without agent**
Bio-degradation	41 variables	Time : 3.23*10^{-4} seconds Memory : 23.92MB	Time : 0.81 seconds Memory : 19,98 MB
Absenteeism	19 variables	Time: 3.67 *10^{-4} seconds. Memory: 25.36 MB.	Time : 0.66 seconds Memory :20.20 MB
Air quality	12 variables	Time : 3.24*10^{-4} seconds Memory :22.54 MB	Time : 0.68 seconds Memory :19.35 MB

TABLE IV. ASSOCIATION RULES RESULT FOR BIO-DEGRADATION DATA

Genetic algorithm for quantitative variables		Association rules phase		
		Association rules with agent	**Association rules without agent**	**Number of rules extracted**
Bio-degradation data	Cluster 1	Time : 0.01 seconds Memory : 23.49 MB	Time : 0.02 seconds Memory : 2.09 MB	80
	Cluster2	Time : 0.01 seconds Memory : 32.01 MB	Time : 0.02 seconds Memory : 18.55 MB	79
	Cluster 3	Time : 0.01 seconds. Memory : 33.67 MB.	Time : 0.02 seconds. Memory : 15.16 MB.	1

TABLE V. ASSOCIATION RULES RESULT FOR AIR QUALITY DATASETS

Genetic algorithm for quantitative variables		Association rules phase		
		Association rules with agent	**Association rules without agent**	**Number of rules extracted**
Air quality data	Cluster 1	Time : 0.01 seconds Memory : 34.99 MB	Time : 0.02 seconds Memory : 25.03 MB	42
	Cluster2	Time : 0.02 seconds Memory : 30.09 MB	Time : 0.02 seconds Memory : 23.49 MB	15
	Cluster 3	Time : 0.01 seconds. Memory : 36.10 MB.	Time : 0.02 seconds. Memory : 23.83 MB.	42

TABLE VI. ASSOCIATION RULES RESULT FOR ABSENTEEISM DATA

Genetic algorithm for quantitative variables		Association rules phase		
		Association rules with agent	**Association rules without agent**	**Number of rules extracted**
Absenteeism data	Cluster 1	Time : 0.02 seconds. Memory : 38.16 MB.	Time : 0.30 seconds. Memory : 13.44 MB.	55
	Cluster2	Time : 0.01 seconds. Memory : 36.72 MB.	Time : 0.02 seconds. Memory : 16.63 MB.	10
	Cluster 3	Time : 0.01 seconds. Memory : 35.86 MB.	Time : 0.02 seconds. Memory : 13.11 MB.	3

Table VII presents a relevant improvement of the global execution time, including clustering and association rules. Concretely, as an average the execution time with the multi-agent integration is 8.33 times less than the traditional approach.

According to the experiments conducted in this section. The integration of multi-agent technology demonstrates the potential of sustaining the process of knowledge extraction from datasets. The results illustrated in Table VI confirm the efficiency of the developed multi-agent framework. However, the consumption of memory used requires an improvement.

The satisfying results obtained in this section with the three varied datasets, make our tool appropriate to solve the issue of agriculture domain covered by this research.

TABLE VII. GLOBAL EXECUTION TIME

Data sets	Total execution time with agent	Total execution time without agent
Bio-degradation	0.08 seconds	0.84 seconds
Air quality	0.09 seconds	0.74 seconds
Absenteeism	0.10 seconds	0.73 seconds

IX. APPLICATION OF THE FRAMEWORK FOR OLIVE CROPS DATASETS

The purpose of this section is mainly the evaluation of the rules returned by the autonomous mechanism of our multi-agent tool. Table VIII shows the eleven variables from the agricultural field in Morocco. Subsections of this part illustrate the results of both clustering and association rules agents.

TABLE VIII. OLIVE OIL VARIABLES

Variables [min, max]		
Fertilization [0, 815]	Picking [1, 998]	Total cost [1.031, 997]
Phyotsanitary treatment [0, 624]	Irrigation [0, 885]	Depreciation cost [0, 769]
Ground work [0, 631]	Direct cost [1.236, 20227]	Olive total cost [0.14, 1.04]
Size [1.25, 625]	Indirect cost [0, 587]	
Total cost/ha [1.729,5072]	Production [730, 12187]	Yield [11.1, 26]
Transformation cost [0.025, 0.04]	Cost of transport [0.004, 0.1]	Olive oil total cost [1.26, 5.75] (variable to predict)

A. Result of Clustering Agent

Table IX shows three clusters generated. The analysis of this result illustrates, an improvement in the memory consumption compared to the three test case datasets. In addition, in this experiment, the clustering agent performs it task with a satisfactory execution time.

B. Result of Association Rules Agent

Table X presents sixty-sixth rules generated. In this experiment particularly in cluster 1, we note that, the execution time of association rules agent is higher compared to that in all test datasets. However, in this phase, there is an improvement in the memory consumption.

C. Interpretation

Table XI, illustrates the rules that satisfy the criteria of minimal olive oil cost. Note that, the rules extracted present high confidence values, which gives more credibility to the results.

From analysing, the rules obtained. Several rules indicate the impact of direct costs on the olive oil cost such fertilization, irrigation, picking and groundwork. In order to improve the value of olive agricultural chain and obtain lower cost of olive oil, the decision makers should focus on the optimization of direct costs.

TABLE IX. RESULT OF CLUSTERING AGENT FOR OLIVE OIL DATASETS

Clusters	**Variables**	**Execution time**	**Memory consumption**
Cluster1	Fertilization, Phytosanitary treatment, Groundwork, size, Irrigation, Indirect cost, Depreciation cost, olive total cost, Yield, Transformation Cost, Cost of transport.	$3.39*10^{-4}$ seconds	14.72 MB
Cluster 2	Picking, direct cost, total cost, Total cost/ha,		
Cluster 3	Production		

TABLE X. ASSOCIATION RULES RESULT FOR OLIVE OIL DATASETS

Clusters	**Number of rules**	**Execution time**	**Memory consumption**
Cluster 1	55	0.29 seconds	28.15 MB
Cluster 2	10	0.03 seconds	24.25 MB
Cluster 3	1	0.02 seconds	32.84 MB

TABLE XI. EXTRACTED RULES

support = 20 (34%) , confidence = 90 % : Fertilization in [10.0; 123.0] --> Olive oil total cost in [1.85; 3.64]
support = 24 (41%) , confidence = 85 % : Size in [14.0; 145.0] --> Olive oil total cost in [1.54; 3.19]
support = 15 (25%) , confidence = 100 % : Ground work in [135.0; 244.0] --> Olive oil total cost in [1.97; 3.74]
support = 11 (18%) , confidence = 78 % : Picking in [336.0; 584.0] --> Transformation cost in [0.03; 0.083] AND Olive oil total cost in [1.94; 3.56]
support = 8 (13%) , confidence = 88 % : Fertilization in [10.0; 50.0] --> Transformation cost in [0.03; 0.083] AND Olive oil total cost in [1.85; 2.71]
support = 11 (18%) , confidence = 91 % Ground work in [10.0; 60.0] --> Transformation cost in [0.036; 0.083] AND Olive oil total cost in [1.6; 2.54]
support = 6 (10%) , confidence = 100 % : irrigation in [183.0; 264.0] --> Transformation cost in [0.025; 0.04] AND Olive oil total cost in [1.85; 2.67]

XI. Conclusion

In addition to the task of analyzing the olive oil, dataset and giving answers about parameters that optimize the cost of olive oil. The key goal of the Moroccan Green Plan is to increase agricultural production and farm income. In order to achieve this goal, public actions should optimize the organization of agricultural value chains and turn to promising sectors such the olive crops.

In the same context, this paper includes results about the factors that improve the value chain of olive oil production. Therefore, in this article, we propose an agent-mining framework that combined two data mining techniques in multi-agent environment. Thus, the use of the proposed framework can be extended into different domain that requires analyzing quantitative datasets.

The multi-agent framework includes data mining algorithms, Weka-based and JADE frameworks and four different types of agents precisely data agent, clustering agent, association rule agent and coordinator agent. The framework was defined to apply different data mining techniques using collaborative approach of interaction among agents to work with an integrated, intelligent perspective that primarily intend to improve the knowledge discovery process.

We tested our framework as potential solution to the issue of optimizing olive oil cost through three test cases from different domains. The results proved that the framework performed well with respect to runtime execution. Thus, the agent-mining tool proves it autonomous aspect and allows extracting sixty-sixth credible rules from olive oil datasets with performed execution time.

In future work, we will study the integration of real time datasets in our multi-agent tool with maintaining reduced runtime processing.

In the same agricultural context, the framework will be adapted to include instantaneous weather parameters and ground composition. Therewith, any upcoming change in the quality of minerals of the soil can be detected by the framework, which could send messages to the biological researchers in order to support preventive actions.

References

[1] ADA, Agricultural Development Agency., “Les Fondements de la stratégie du Plan Maroc Vert.” 2014.

[2] Faysse, N., and C. Simon. “Holding All the Cards Quality Management by Cooperatives in a Moroccan Dairy Value Chain.”European Journal of Development Research, 2014.

[3] Obidallah, W.J., Raahemi, B. & Ruhi, U, “Clustering and Association Rules for Web Service Discovery and Recommendation: A Systematic Literature Review”. SN COMPUT. SCI. 1**,** 27, 2020.

[4] University Waikato. Weka machine learning project, September 2009.

[5] R. Agrawal, T. Imielinski, and A. N. Swami, “Mining association rules between sets of items in large databases”. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 1993.

[6] F. Moslehi and A. Haeri, “A genetic algorithm-based framework for mining quantitative association rules without specifying minimum support and minimum confidence”, Scientia Iranica D 27(3), pp.1316-1332, 2020.

[7] Gonzales, E., Mabu, S., Taboada, K., Shimada, K., Hirasawa, K., “Mining Multi-class Datasets using Genetic Relation Algorithm for Rule Reduction”. In IEEE Congress on Evolutionary Computation, CEC 2009, pp. 3249–3255, 2009.

[8] Shah, S., Rashid, M. and Arif, M, “Estimating WCET using prediction models to compute fitness function of a genetic algorithm.” *Real-Time* Syst 56**,** pp**.**28–63, 2020.

[9] Zayrit Soumaya, Belhoussine Drissi Taoufiq, Nsiri Benayad, Korkmaz Yunus, Ammoumou Abdelkrim,,”The detection of Parkinson disease using the genetic algorithm and SVM classifier”, Applied Acoustics,Volume 171, 2021.

[10] Mesbahi, N., Kazar, O., Benharzallah, S., Zoubeidi, M.,”A Cooperative Multi-Agent Approach-Based Clustering in Enterprise Resource Planning”. International Journal of Knowledge and Systems Science, 6(1) pp. 34-45, 2015.

[11] Gelvez Garcia, Nancy Yaneth; Balenduarte, Andrés David, Espitia Cuchango and Helbert Eduardo.” Multi-Agent System Used for Recommendation of Historical and Cultural Memories”. Tecciencia, vol.14, n.26, 2019.

[12] Calegari, R., Ciatto, G., Mascardi, V. *et al.*,”Logic-based technologies for multi-agent systems: a systematic literature review”. Auton Agent Multi-Agent Syst 35**,** 1, 2021.

[13] Qian A., Chen CY.,” Accurate Model of the Internet Financial Poverty Alleviation Based on the Multi-agent and the Data Mining”. In Data Processing Techniques and Applications for Cyber-Physical Systems, DPTA 2019. Advances in Intelligent Systems and Computing, vol 1088. Springer, 2019.

[14] Popa, H.,Pop, D., Negru,V., Zahariae,D. ,” Agent Discover : A Multi Agent System for knowledge discovery”. In IEE/WIC/ACM International conference on Web Intelligence and Intelligent Agent Technology, Sydney,pp. 571-574 , 2008.

[15] Tong, C., Sharma, D., Shadabi, F.” A Multi-Agents Approach to Knowledge Discovery”. In IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Sydney, pp. 571–574, 2008.

[16] Nahar, J., Imam, T., Tickle, K., Chen, Y. “Association rule mining to detect factors which contribute to heart disease in males and females.”, Expert Systems with Applications 40(4), pp. 1086–1093, 2013.

[17] Ait-Mlouk A., Agouti T., Gharnati F.”Multi-agent-based modeling for extracting relevant association rules using a multi-criteria analysis approach”, Vietnam J Comput Sci 3:235–245, 2016.

[18] Kaur, H., Chauhan, R., Alam, M. A,.Aljunid, S., Salleh, M.: SpaGRID, “A special Grid Framework for High Dimentional Medical Databases”. In HAIS 2012, Part 1. LNCS, vol. 7208, pp.690-704, Springer, Heidelberg, 2012.

[19] F. Bellifemine, A. Poggi, and G. Rimassi. JADE,”A FIPA-Compliant agent framework”, Proceedings Practical Applications of Intelligent Agents and Multi-Agents, 1999.

[20] Dua, D. and Graff, C., “UCI Machine Learning Repository”, CA: University of California, School of Information and Computer Science, Irvine, 2019.

[21] A.Salleb-Aoussi, C.Vrain and C. Nortet, “Quantminer, a Genetic Algorithm for Mining Quantitative Association Rules “, IJCAI-07 1036, 2013.

[22] I.Belabed, M.Talibi Alaoui and A.Belabed, “Association Rules Algorithms for Data Mining Process Based on Multi Agent System”. MLN 2019, LNCS 12081, pp. 431–443, 2020.

Development of Smart Healthcare System for Visually Impaired using Speech Recognition

Smart Healthcare System

Khloud Almutairi[1], Samir Abdlerazek[3]
Hazem Elbakry[4]
Information Systems Dept. Faculty of Computers and Info.
Mansoura University, Mansoura, Egypt

Ahmed Ismail Ebada[2]
Research and Development Dept.
GlaxyTech, Munich, Germany

***Abstract*—This paper presents a solution for the Visually Impaired (VI) based on wearable devices. VI people need support or a guide to support them to locomote from one place to another. Using Wearables help users to achieve a great understanding of the surrounding environment. The proposed system is based on wearable smart glasses to support VI to locomote. It provides a solution integrated with speech recognition to get the destination name and look for the routes. The proposed system is based on Google maps with speech recognition to work as user assistance. The results of the research results proved that the system works with high accuracy of 99% and can help the person as an effective tool for localization guidance. The system can assist VI people to move and have a better life quality.**

Keywords—Personal assistance; speech recognition; visually impaired person assistant; smart wearable device; smart sensory system

I. Introduction

There are about 285 million people who are visually impaired (VI) worldwide based on World Health Organization (WHO). From them, there are about 246 million have low vision and 39 million are totally blind. About 90% of the world’s visually impaired live in low-income conditions [1]. Most of VI depend on other people to do their daily activities [2]. This problem is one of the challenges in Europe and USA, because most of VI and disabled live by themselves [3]. One of the biggest obstacles for visually impaired people is that feeling figure. Adapting to experience failure is difficult and takes time. Some people see that the supervisor or support group will help them learn to accept their experience loss so they can move forward toward a richer experience. The significant factor in this often involves developing the skills required to live independently [4]. People who are visually impaired may be born with experience failure or create a visual disability later in life as the result of the accident or heart illness. Some other statements are used to describe the varying degrees of experience loss the individual may have. The terms “visually impaired” and “visual disability” are used to allow all people with reduced experience, irrespective of the severity of experience failure or blindness. However, the following blindness statements and descriptions offer a better explanation of the individual's practical experience [5].

One of the greatest applications she utilizes for mobility is Blind Square, the self-voicing GPS app for those visually impaired that announces crossings and levels of curiosity. Blind Square works with Foursquare, then favorite restaurants, cafés, shops, and other jobs that user frequently go into are flagged. That gives blind users an unbelievable amount of independence, as they don’t gets to take fellow pedestrians which stores or landmarks they are returning. Blind Square gets them learn [6].

This paper presents a smart wearable sensors system with speech recognition to assist VI people to locomote safely and independently. The proposed solutions employed sensors to track the distance in front of the VI user, alert them, and keep their family the possibility to track them or assist them.

II. Related Work

There are many researchers used accelerometers and vibration feedback based on gesture recognition as a successful simple user interaction system [5]. A. Ismail et al [6] proposed a healthcare system based on a speech recognition system for smart homes and smart hospitals to help disabled people to control the appliances and depend on themselves on daily based activities.

In [7-19] authors proposed different solutions by integrating speech recognition solutions, a Global positioning system (GPs), and the Internet of Things (IoT) to provide VI with a simple solution that can help them to understand the surrounding area.

Recently, there are concentrations of research on using speech recognition applications like Amazon Alexa in chat bot's application as the interaction between users and devices to find the places, services, and trigger actions [20]. The paper presents a solution that may help visually impaired people to locomote using speech recognition and Google API. We used the research from T. Ashwell et al. [21] as a base to build a speech recognition system for visually impaired people. They used the ASR system to recognize the speech commands from Japanese students who learn English.

In our proposed research, the speech recognition system with the Google API is used to recognize a command from an English speaker based only on some predefined locations by the user. The target from recognition using Google API to find the previously saved locations with a speech command and get the direction path descriptions as a verbal voice. The system helps the user with localization and with object detection to reach the destination. Besides that, it can detect the obstacle that VI cannot see and the system can ask for video assistance with a caregiver on-demand when it is necessary.

III. Proposed Smart Glasses

A. Proposed System Features

The proposed smart glasses track the street in front of the VI user up to 400 cm and it is worn on the user's eyes. The system works as an alert system that uses an ultrasonic sensor and camera for object detection ahead of the user. The system helps users who wear smart glasses to avoid stumble overdue to the alarm system which is based on the obstacle detection system. The obstacle detection system assists VI users to get alarms using visual recognition and ultrasonic sensors. The caregivers of VI users play only a role when they get a help request from the VI user using a speech recognition command “help” or when a free fall is detected, so they can start monitoring the user and watch a streaming video from the installed camera.

The proposed smart glasses provide a smart solution for VI users to locomote using visual recognition using a camera, a speech recognition system to help the user to provide the system with the target destination. The system uses a sound guide using Google APIs to provide the user with the instruction. It supports 4G connections using a cellular module and GPS modules to find the routes. The proposed system is on sleep mode until it detects a speech command from the user. The proposed smart glasses are lightweight and inexpensive.

B. Architecture

The smart glasses use a rechargeable battery as the power source. As shown in Fig. 1, the system structure is based on sensors as inputs, a controller (Raspberry Pi), and outputs as alerts.

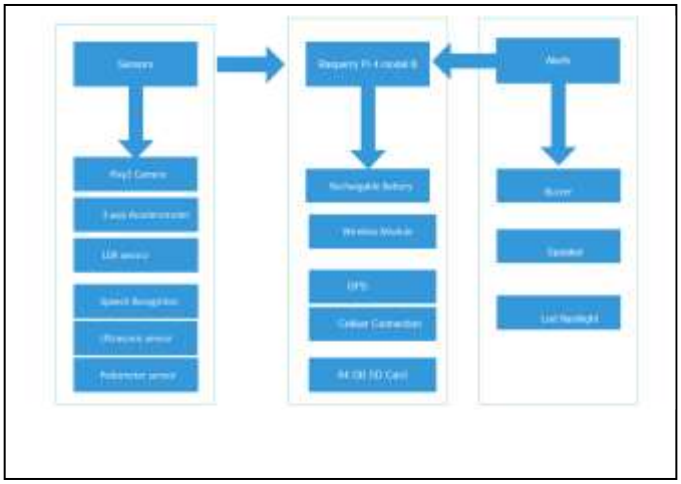


Fig. 1. The Proposed System Block Diagram.

The system used multiple sensors on the proposed smart glasses to work as a guide system for VI users. The system provided two different web APIs on the cloud one for caregivers and the second one for VI users as shown in Fig. 2. The caregiver API accepts provide them with a view of alerts, a possibility to monitor VI user, and a sending voice instruction. The VI user can add new destinations, add new caregivers, ask for help, and receive instructions. The sensors are connected to Raspberry Pi (Fig. 3) which sends the alerts and connects the smart glasses to the internet using the GSM module. The data from smart glasses are saved on the user profile on the cloud.

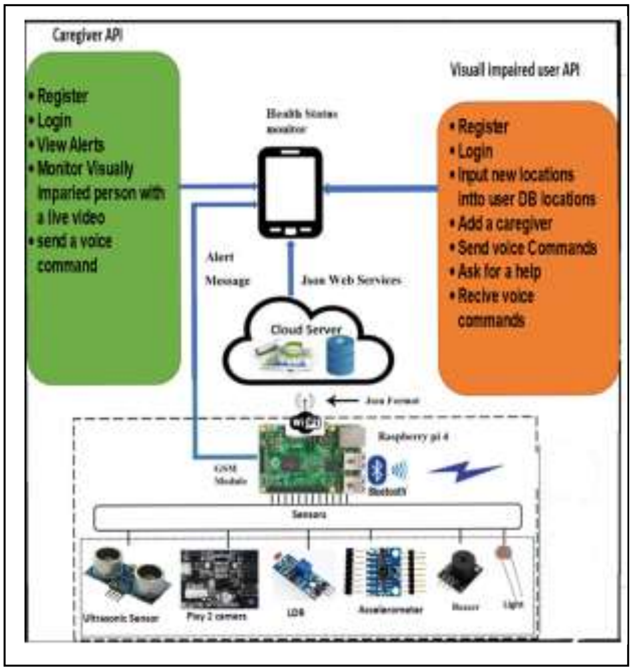


Fig. 2. The Proposed System Architecture.

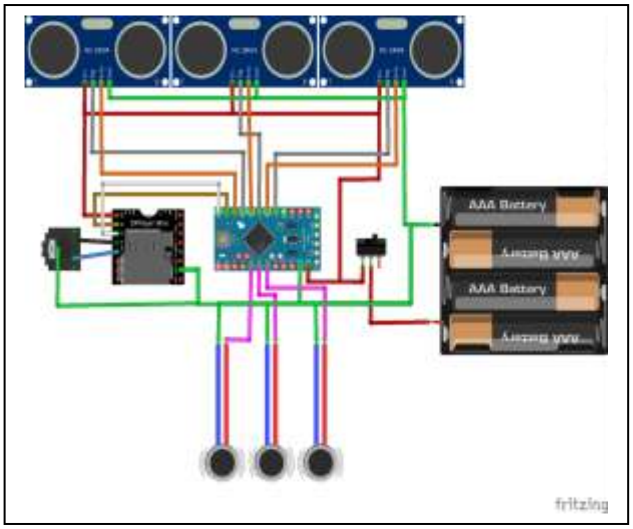


Fig. 3. The Sensor's Connections with the Raspberry Pi Board.

1) The system controller: The system is based on the Raspberry pi4 board (Fig. 4). This board has a 64-bit quad-core processor, it is possible to have RAM up to 4GB, USB 3.0, dual-band 2.4/5.0 GHz wireless LAN, Bluetooth 5.0, and Gigabit Ethernet. The board is used rather than Arduino because the proposed system needs a powerful system to handle speech recognition and visual recognition processing.

2) Sensors: The proposed smart glasses contain different types of sensors to work as a guide tool. They contain an ultrasound sensor, camera, accelerometer, LDR sensor, and GPS. The overall system depends on the interaction between the different multi-sensors signals to decrease the false alarms as shown in the block diagram in Fig. 5.

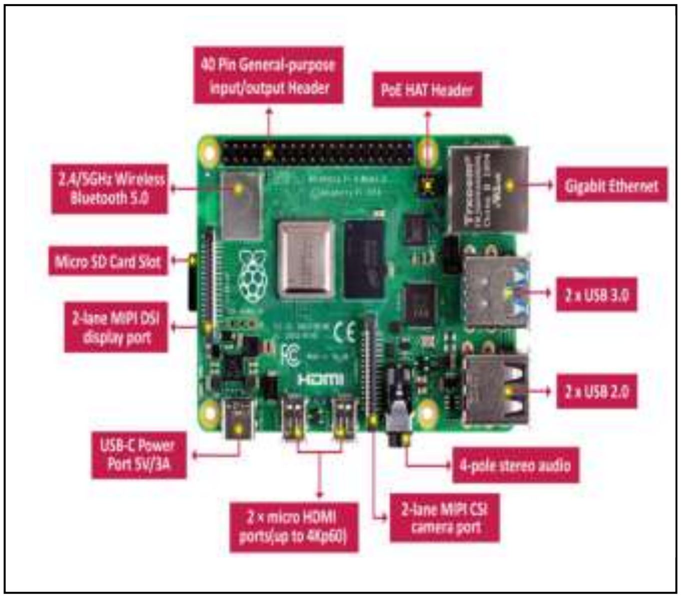

40 Pin General-purpose input/output Header
PoE HAT Header
2.4/5GHz Wireless Bluetooth 5.0
Gigabit Ethernet
Micro SD Card Slot
2 x USB 3.0
2-lane MIPI DSI display port
2 x USB 2.0
USB-C Power Port 5V/3A
4-pole stereo audio
2 x micro HDMI ports(up to 4Kp60)
2-lane MIPI CSI camera port

Fig. 4. Raspberry Pi Model 2.

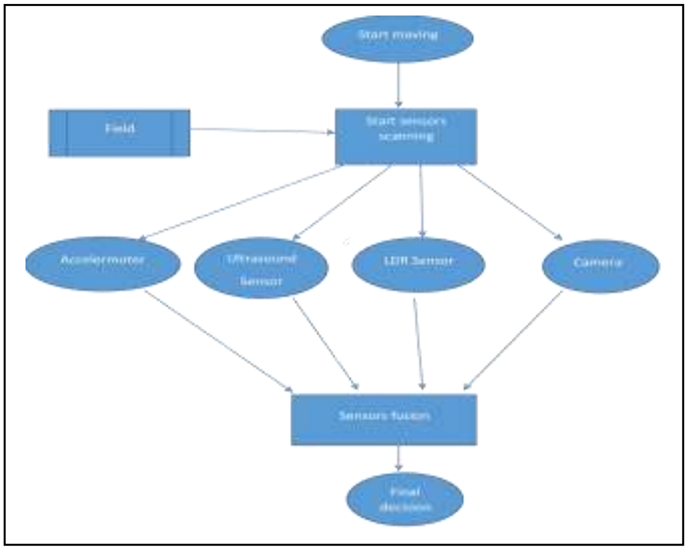


Fig. 5. The Sensory Decision Support System.

The proposed system components are detailed in the following:

a) Ultrasound Sensor (US): The proposed system employs US to detect the obstacles in front of the VI user. USs detect the distance using ultrasonic waves as shown in Fig. 6. The sensor head emits an ultrasonic wave that is reflected from the target object. Ultrasonic sensors detect the distance to the target object by measuring the time between sending and receiving the ultrasonic wave. An optical sensor has a transmitter and a receiver, while an ultrasonic sensor uses the same ultrasonic element to transmit and receive. In a reflection type ultrasonic sensor, the ultrasonic waves are alternately transmitted and received by a single oscillator. This means that the sensor head can be kept very small. The distance can be calculated using the following formula:

Distance $L = 1/2 \times T \times C$

where L is the distance, T is the time between transmission and reception, and C is the speed of sound (the value is multiplied by 1/2 because T is the time it takes the wave to travel there and back).

The ultrasound sensor sends the sum of the distance between the sensor and the nearest obstacle and the distance can be calculated by dividing the total distance by 2 because the total distance is the distance from the sensor up to the obstacle forward and backward. The sensor sends the distances in centimeters when it is connected to the Arduino board as shown in Fig. 7.

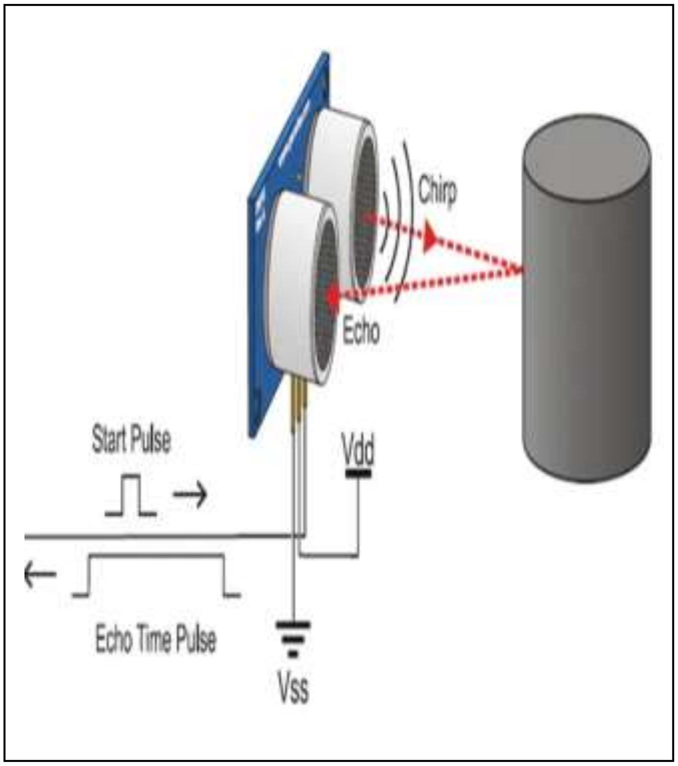
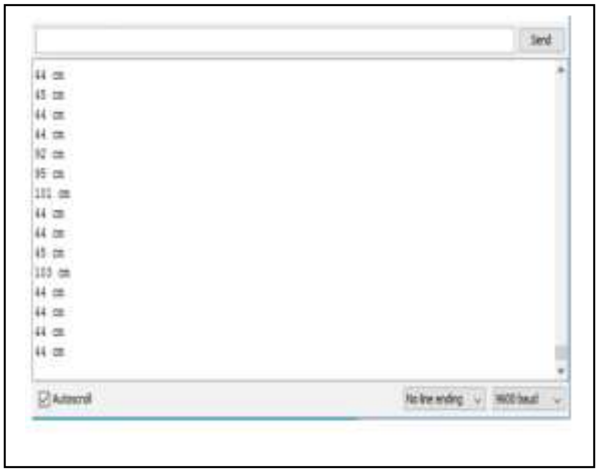


Fig. 6. The Ultrasound Sensor.



```
44 cm
45 cm
44 cm
44 cm
92 cm
95 cm
101 cm
44 cm
44 cm
45 cm
103 cm
44 cm
44 cm
44 cm
44 cm
```

☑ Autoscroll   No line ending   9600 baud

Fig. 7. The Measurements of the Ultrasound Sensor.

b) Camera: The proposed system is provided with a Pixy2 Camera (Fig. 8) which has two functionalities. The first is to detect obstacles by the recognition of the obstacles and recognizing the streets to guide VI users as a human eye because GPS has some errors in defining the exact location. The second function of the camera is to work as a watching person to provide the caregiver a possibility to watch the VI user on demand. The Pixy 2 camera is a great solution for visual recognition applications because it is easy to train the camera to recognize the objects. The camera is used to help VI user to recognize when there is a hole, Lighting poles, or dangerous places.

c) LDR Sensor: LDR sensor is used to detect darkness to turn a led on the smart sunglasses to provide a clear video streaming for caregivers when their help is needed by the VI users. As shown in Fig. 9, the led is turned on only when the LDR detects the darkness and then the circuit for the led will be closed to turn it on.

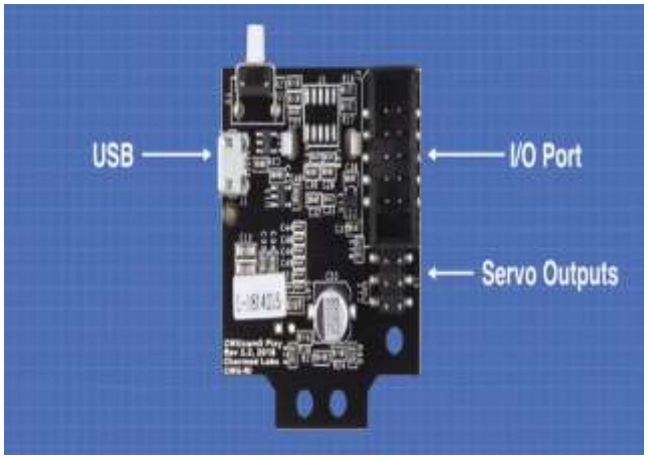


Fig. 8. Pixy2 Camera for Image Recognition.

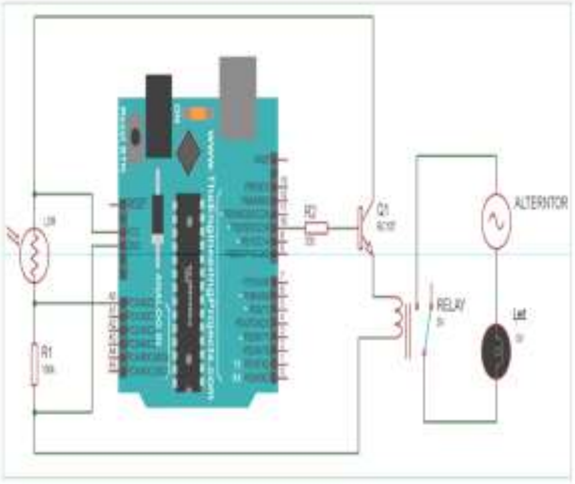


Fig. 9. The Led Connection with LDR Sensor

C. Speech Recognition

The paper presents smart glasses with a speech recognition module, which is trained to recognize the speech of the user. As shown in Fig. 10 the user starts to give an input speech as the name or location name. Then applying feature extraction to find a similarity score for the command based on a database of the speaker model with the user-defined commands.

The proposed system used the speech recognition module in Fig. 11 installed on the smart glasses. This module is a compact and easy-to-use speech recognition board. It is a loudspeaker dependent speech recognition module. It supports up to 80 voice commands in total. A maximum of 7 voice commands can work at the same time. Each tone can be trained as a command. The users must first train the module before they can recognize any voice command. It has 2 control options: serial port (full function), general input pins (part of the function). Common output pins on the board can generate different types of waves while recognizing appropriate voice commands. The V2 speech recognition module supports 15 commands in all and only 5 commands at the same time. In V2, voice commands are divided into three groups as you practice them. And only one group (5 commands) could be imported into Recognizer. It means that only 5 voice commands are effective at the same time. On V3, voice commands are stored in a large group like a library. All 7 voice commands in the library can be imported into the recorder. Parameters: Voltage: 4.5 - 5.5 V, current: <40 mA, digital interface: 5 V TTL level for UART interface and GPIO, analog interface: 3.5 mm mono channel microphone connection + microphone pin interface. Dimensions: 31 x 50 mm. Detection accuracy: 99% (in ideal surroundings). Features: Supports a maximum of 80 voice commands, with each voice 1500 ms (speak one or two words). A maximum of 7 voice commands effectively simultaneously. Arduino library is included. Easy control: UART / GPIO. User-controlled general pin output.

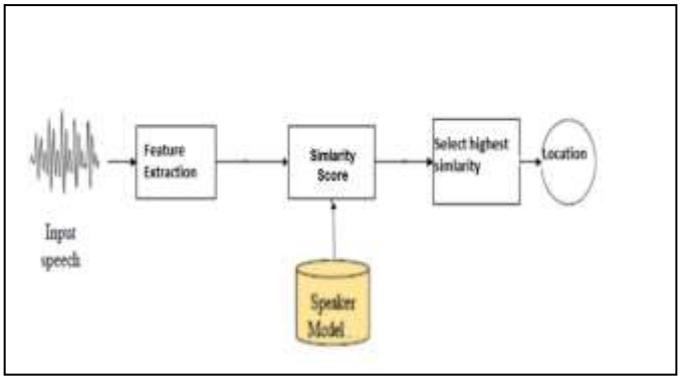


Fig. 10. System Overview.

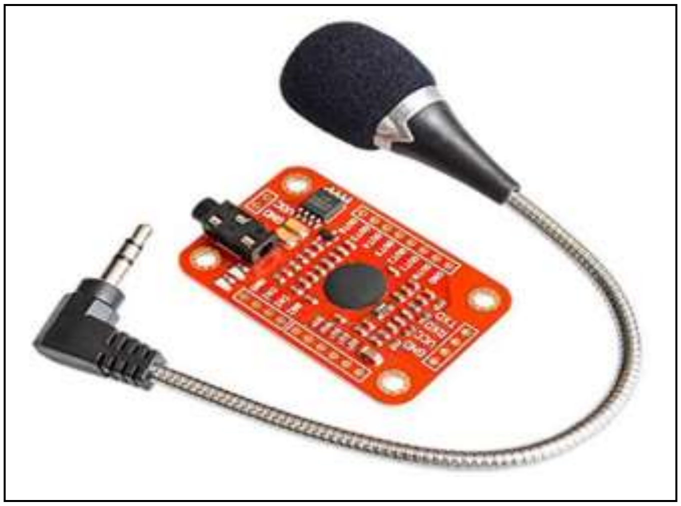


Fig. 11. The Speech Recognition Module.

The proposed system as shown in Fig. 12 starts with the recognition process by matching the input command against the command templates by using the Dynamic Time Warping (DTW) algorithm which is based on measuring similarity between two temporal sequences. The DTW is used for authentication purpose if the user is authorized, then the system stat to receive the command which contains the destination name as pre-recorded in the speaker database. If it is found in the database, then the user can get a description of the route to the destination.

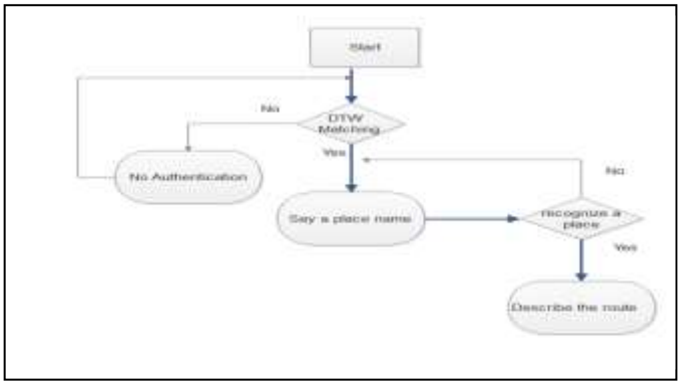


Fig. 12. The Proposed System Flowchart.

The system is based on Google Map APIS to get the route description to describe the route.

D. Visual Recognition

The camera in the proposed system is used for two purposes, first as an object recognition tool to detect an obstacle in front of the VI user or as an on-demand streaming tool to show the path to the caregivers to describe the path for the VI people as shown in Fig. 13.

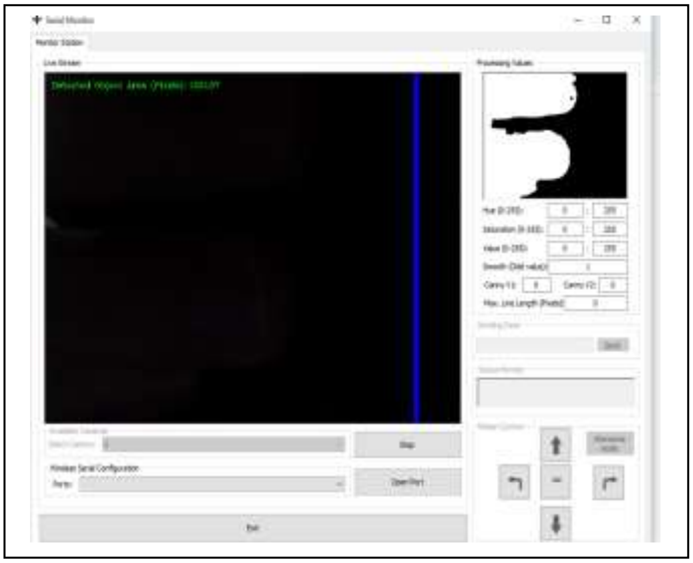


Fig. 13. The Camera Interface on the Caregiver Side.

E. System Implementation

The proposed system is comprised of Raspberry Pi and multiple sensors. The software is compiled on the board. Every sensor has a code to send the parameters to the board. The Raspberry pi board and the sensors are integrated with the open-source code community. Functions such as communications and GPS readings are available as libraries, which simplify the code and implementation. The proposed system is supported by a battery for the components and the controller. The battery can work for 2 days and the user can extend the battery time with a power bank if it is needed. The sensors, which were connected to the board, send signals that trigger the predefined events. The ultrasonic sensor gets the power directly from the battery because it is active most of the time while the system is running.

The board requires 5.5 V and has a low power consumption. The alarms, including sound, and camera to transmit video to the caregivers are triggered by the sensor events and therefore consume power during short periods. The proposed system proposed a complete coverage algorithm to define the path based on Google maps. The system uses an obstacle avoidance algorithm to cross through obstacles. The pseudo code of this algorithm is:

Pseudo code for the user's motion

```
if (the camera recognized an obstacle && Ultrasonic sensor detected an obstacle)
{
        Stop
        send an alert signal
                        }
    }
        if (the accelerometer detected a free fall of the user)
        {
            If the Pedometer did detect motion for 30 seconds
            {
                Alert
                Send a call for a help
            }
        }
}
```

If the sum of all probabilities of overall the system is more than one, the systems give a sound command with the direction and the exact location of the obstacle to overcome it. The aim of this algorithm to detect the obstacles and then check whether they are high, or not. If the camera detected an obstacle that is a danger sign or any predefined object, it sends a signal to the caregiver as shown in Fig. 14 and 15.

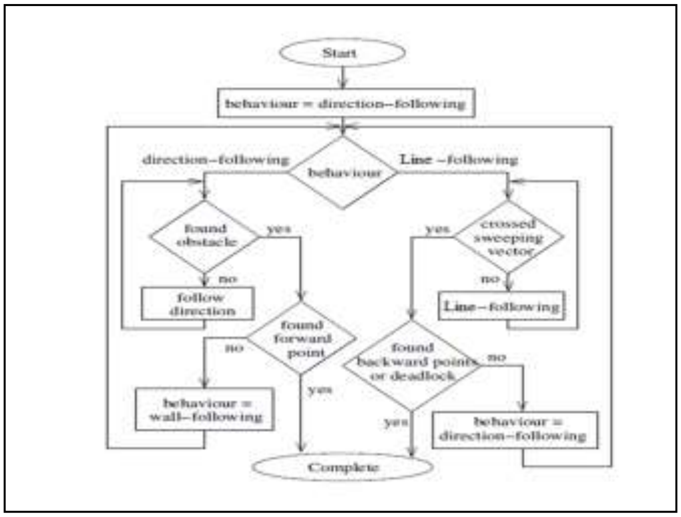


Fig. 14. The Obstacle Avoidance Flowchart.

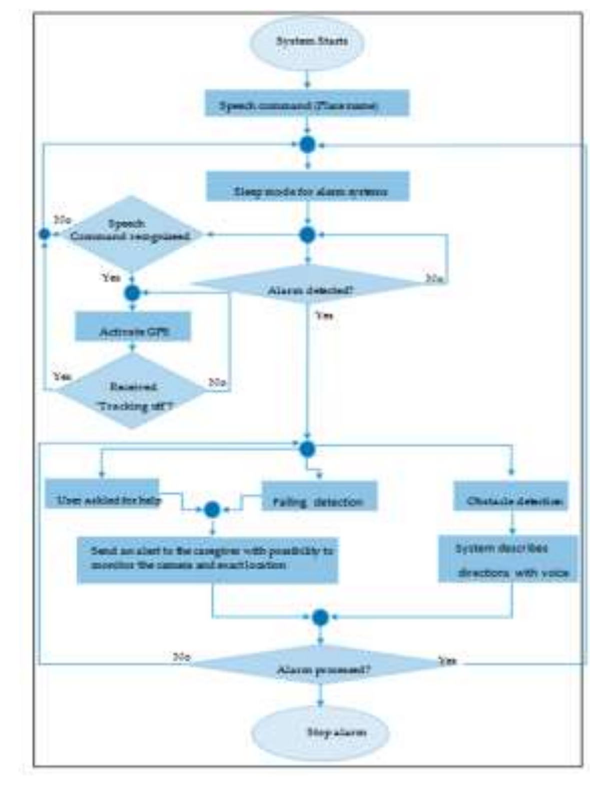

System Starts
Speech command (Place name)
Sleep mode for alarm systems
No
Speech Command recognized
Yes
Alarm detected?
No
Yes
Activate GPS
Yes
Received 'Tracking off'?
No
User asked for help
Falling detection
Obstacle detection
Send an alert to the caregiver with possibility to monitor the camera and exact location
System describes directions with voice
No
Alarm processed?
Yes
Stop alarm

Fig. 15. Basic Flowchart of the System Operation.

If(P(S1) +P(S2) +P(S3) ≥1

Then: There is a mine

P (S1) is sensor probability of obstacle existence by ultrasonic sensors.

P (S2) is a sensor probability of darkness existence by the LDR sensor.

P (S3) is a sensor probability of danger existence by the camera.

The proposed smart glasses are as in Fig. 16, a light tool that works like a small computer for visual recognition, speech recognition, and an interaction tool between VI users and their caregivers.

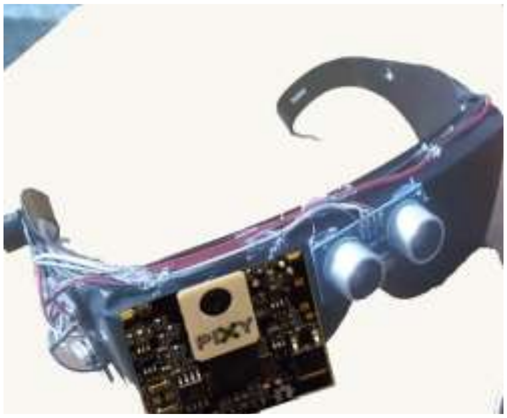


Fig. 16. The Smart Glasses Prototype.

IV. Tests and Results

The proposed system is tested by users with different accents with an age range between 15 and 30. The test is based on speech commands and locomotion. The proposed system showed efficiency and adaptability with different situations because it gives the user possibility to add new commands to new destinations, new objects to be recognized by the camera. The proposed system was very comfortable for the user while testing because they need to wear only glasses. The system has 15 commands for testing the first section of usually used items designed to elicit performance on 15 grammatical features (possessives, plural -s, 3rd person -s, questions, comparative adjectives, relative clauses, conditionals, modal verbs, relative adverbs, verbs. Multiple instances of each feature appeared in the test. Items ranged in length between 2 and 4 words. The 15 items in each section were arranged in random length order so that the items would be not relatively to find out the performance of the recognition according to a limited number of matching items.

The test items are formulated as orders that contain in between one of the preserved entries as target destinations. The user must record the commands as the name of the destination as a separate entry. The entries as shown in Table I should be short and has a defined name with the user speech and a saved spot on the Google map account. The proposed application is designed and developed with the Xamarin tool; it is a tool from Microsoft to develop cross-platform applications. The application can work on android, windows, and iOS platforms. The application is developed as simple as possible to be used easily by visually impaired people. The application just expects a word of some words from the user to start looking for the target destinations.

TABLE I. Used Sentences in the System Test

The Sentence	The order
Go Home	Home
Go to my Work	Work
Go to Metro station	Metro
Go to the train station	Train
Go to Restaurant	Restaurant
Go my friend Ali	Ali
Go my sister Lana	Lana
Entertainment near to me	Entertainment
I go now	Green
Call police	Police
Call Ambulance	Ambulance
Go to a concert	Concert
Go to the nearest Supermarket	Market
Call emergency	Call Emergency
Where am I	Where

After testing the system with those registered 15 sentences, the system worked with a success rate of 99%. We tested even changing the order of the sentences with the system with native speaker input to test the system accuracy and the system could match them successfully. The best practice with the speech orders to try to use short sentences with a place or an alias name of the destinations or friends' names. The use of a camera mounted on the smart glasses provided the system with real-time images from the environment to provide the caregivers with the possibility to help VI people in difficult situations.

V. Discussion

The proposed system proposed a solution with speech recognition and visual object identification to help VI people in navigation. We developed a system that improved the proposed system from RAMADHAN [18] by improving the accuracy of speech recognition and we added more details of the detection of the objects in front of the VI person by using video recognition by an installed camera on the smart glasses. The system gives the VI people a better experience because it is lightweight smart glasses that is more convenient to wear than a wrist wearable device. The system provides a tracking system for the VI person that can be accessed by a family person to lead them in hard situations. There were some challenges while developing the system such as the design size and battery life. The system provided an optimal low-cost solution with most of the features that can help the VI to navigate by themselves.

VI. Conclusion

The paper presented a successful solution for VI users to help them to overcome the obstacle they face while the walk alone from point A to point B. The proposed system in this paper is smart glasses that contain a Raspberry Pi board, camera, ultrasound sensor, speech recognition module, and LIDR sensor. The proposed system provided a successful solution based on Google Maps APIs to describe the path to the users based on the pre-recorded commands of the destination the VI user usually visit. The system proposed an applicable solution based on machine learning and IoT [22-27] that can help VI people to have a better life and locomote.

Acknowledgment

- The article authors declare there is no interest conflicts in the article.
- Consent for publication: Informed consent was obtained from all individual participants included in the study.
- Data availability: Data is available on request.
- The article did not get funds for the research.
- We accept all ethics approval and consent to participate in the scientific research area.
- All Authors worked on the proposed research.
- Acknowledgments: We would like to thank GlaxyTech, Munich, Germany for supporting devices and sensors and all needed requirements for implementations.

REFERENCES

[1] R. Jothi, M. Kayalvizhi, K. Sagadevan, “Smart walking stick for visually challenged people”. Asian J. Appl. Sci. Technol. 2017, 1, 274–276.

[2] World Health Organization. Available online: http://www.who.int/en/ (accessed on 9 October 2020).

[3] T. Volkel, R. Kuhn, G. Weber, "Mobility Impaired Pedestrians Are Not Cars: Requirements for the Annotation of Geographical Data. In Miesenberger", Springer, Heidelberg, vol. 5105, pp. 1085–1092, 2008.

[4] Sangami, A.; Kavithra, M.; Rubina, K.; Sivaprakasam, S. “Obstacle detection, and location finding for blind people”. Int. J. Innov. Res. Comput. Commun. Eng. 2015, 3, 119–123.

[5] Kher Chaitrali, S.; Dabhade Yogita, A.; Kadam Snehal, K.; Dhamdhere Swati, D.; Deshpande Aarti, V. An intelligent walking stick for the blind. Int. J. Eng. Res. Gen. Sci. 2015, 3, 1057–1062.

[6] Ismail, A., Abdlerazek, S., & El-Henawy, I. M. (2020). “Development of Smart Healthcare System Based on Speech Recognition Using Support Vector Machine and Dynamic Time Warping”, Sustainability, 12(6), 2403.

[7] M. Aziz, T. Owens, U. Zaman, “Received Signal Strength Indicator-Based Localization of Bluetooth Devices Using Trilateration: An Improved Method for the Visually Impaired People”, International Journal of Electronics and Communication Engineering, 2019.

[8] N. Ramaprasad, P. Narayanan, “Volunteered Geographic Information System and Its Contribution in Service Sector Employment”, In Geographic Information Systems. IntechOpen, 2019.

[9] S. Kotyan, N. Kumar, P. Sahu, V. Udutalapally, “Drishtikon: An advanced navigational aid system for visually impaired people”, arXiv preprint arXiv, pp. 19040.10351, 2019.

[10] A. Montuwy, B. Cahour, A. Dommes, “Using Sensory Wearable Devices to Navigate the City: Effectiveness and User Experience in Older Pedestrians”, Multimodal Technologies and Interaction, vol 3(1), 2019.

[11] Khatri, A. Assistive vision for the blind. Int. J. Eng. Sci. Invent. 2014, 3, 16–19.

[12] Gayathri, G.; Vishnupriya, M.; Nandhini, R.; Banupriya, M.M, “Smart walking stick for visually impaired”,Int. J. Eng. Comput. Sci. 2014, 3, 4057–4061.

[13] Rao, B.; Deepa, K.; Prasanth, H.; Vivek, S.; Kumar, S.N.; Rajendhiran, A.; Saravana, J. Indoor navigation system for the visually impaired person using GPS. Int. J. Adv. Eng. Technol. 2012, 3, 40–43.

[14] Chandana, K.; Hemantha, G.R. Navigation for the blind using GPS along with portable camera-based real-time monitoring. SSRG Int. J. Electron. Commun. Eng. 2014, 1, 46–50.

[15] Kumar, M.N.; Usha, K. Voice-based guidance and location indication system for the blind using GSM, GPS, and optical device indicator. Int. J. Eng. Trends Technol. 2013, 4, 3083–3085.

[16] Q. Rabbani, G. Milsap, N. Crone, “The potential for a speech brain-computer interface using chronic electrocorticography”, Neurotherapeutics, vol 16(1), pp. 144-165, 2019.

[17] José, J.; Farrajota, M.; Rodrigues, J.M.; du Buf, J.M.H. The SmartVision local navigation aid for blind and visually impaired persons. Int. J. Digit. Content Technol. Appl. 2011, 5, 362–375.

[18] RAMADHAN, A. Wearable smart system for visually impaired people. Sensors, 843, 2018, 18.3.

[19] N. Giudice, W. Whalen, T. Riehle, S. Anderson, S. Doore, “Evaluation of an Accessible, Real-Time, and Infrastructure-Free Indoor Navigation System by Users Who Are Blind in the Mall of America”, Journal of Visual Impairment & Blindness, 2019.

[20] A. Nassif, I. Shahin, I. Attila, M. Azzeh, K. Shaalan, “Speech Recognition Using Deep Neural Networks: A Systematic Review”, IEEE Access, vol 7, pp. 19143-19165, 2019.

[21] T. Ashwell, J. Elam, “How Accurately Can the Google Web Speech API Recognize and Transcribe Japanese L2 English Learners” Oral Production? Jalt Call Journal, vol 13(1), pp. 59-76, 2017.

[22] Ahmed Ismail, Samir Abdlerazek & Ibrahim M. El-Henawy, “Applying Cloud-Based Machine Learning on Biosensors Streaming Data for Health Status Prediction”, The 11th International Conference on Information, Intelligence, Systems and Applications 15 – 17 July 2020 Piraeus, Greece, IEEE.

[23] Ahmed Ismail, Samir Abdlerazek & Ibrahim M. El-Henawy. "BIG DATA ANALYTICS IN HEART DISEASES PREDICTION." Journal of Theoretical and Applied Information Technology 98, no. 11 (2020).

[24] A. Ismail, A. Shehab, I. El-Henawy, " Healthcare Analysis in Smart Big Data Analytics: Reviews, Challenges, and Recommendations", In Proceedings of Springer, 2018.

[25] A. Ismail, A. Shehab, I. El-Henawy, "Quantified Self Using IoT Wearable Devices", In the International Conference on Advanced Intelligent Systems and Informatics 2017. AISI 2017. Advances in Intelligent Systems and Computing, vol 639. Springer, 2017.

[26] A. Ismail, M. Elmogy, H. ElBakry, "Landmines Detection Using Low-Cost Multisensory Mobile Robot", Journal of Convergence Information Technology, Volume 3, Issue 4, November 2015.

[27] A. Ismail, M. Elmogy, H. ElBakry, " Landmines Detection Using Autonomous Robots: A Survey", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 4, 4, July-August 2014.

Medical Data Classification using Fuzzy Min Max Neural Network Preceded by Feature Selection through Moth Flame Optimization

Ashish Kumar Dehariya[1]
Computer Engineering Department, Institute of Engineering and Technology
Devi Ahilya Vishwavidyalaya, Indore, India

Pragya Shukla[2]
Professor, Computer Engineering Department, Institute of Engineering and Technology
Devi Ahilya Vishwavidyalaya, Indore, India

***Abstract*—Prediction of the diseases are possible using medical diagnosis system. This type of health care model can be developed using soft computing techniques. Hybrid approaches of data classification and optimization algorithm increases data classification accuracy. This research proposed applications of Moth Flame optimization (MFO) and Fuzzy Min Max Neural Network (FMMNN) for the development of medical data classification system. Here MFO algorithm considers bulk of features from the disease dataset and produces optimized set of features based on fitness function. MFO is able to avoid local minima problem and this is the main cause behind production of optimal set of features. Optimized features are then passed to FMMNN for classification of malignant and benign cases. As classification is concerned, model experiment achieved 97.74% accuracy for Liver Disorders and 86.95 % accuracy for Pima Indian Diabetes dataset. Improving the medical data classification accuracy is directly related to attain good human health.**

***Keywords*—Moth flame optimization; nature inspired optimization; feature selection; fitness function; fuzzy min-max neural network**

I. Introduction

Nowadays, specialized computer software is very popular for medical data classification. For the development of smart healthcare system in smart cities use medical diagnosis software plays a vital role. In this type of medical data classification tool/software, patient’s disease symptoms are generally learned then this learned knowledge is used for deciding the facts whether patients are suffering from disease or not. Here diagnosis of the disease is performed by applying soft computational techniques for classifying the disease datasets. Medical data classification tool needed and demanded in medical field because it overcomes the problem like difficulties in huge medical data analysis by medical professionals means they hardly processes the large amount of medical data, sometimes they may suffer with fatigue or other tensions resulting classification error. So they need a tool/ expert software that can help them to make a justified decision. Aim of our proposed model is to develop a robust medical data classification system for the numerical type medical dataset.

In our model we are using Moth Flame Optimization (MFO) for feature selection and Fuzzy Min Max Neural Network (FMMNN) for Classification of disease datasets. MFO algorithm can be used for selecting best features hence it can be used with classification algorithms to get efficient result [9]. MFO is a genetic algorithm which is inspired from nature. It discovers optimal features to be supplied to data sets classifier. It simulates the behavior of the moths moving towards light source [20]. FMMNN is a supervised classifier. It is inspired from neural structure of the human brain and capable to learn data, accordingly conclude intelligent decisions. FMMNN is a bundle comprising fuzzy logic and min max neural network. FMMNN consist three layers. First (Input) layer contains the number of input nodes which accept input features of supplied datasets. Based on the input features, Different logically separated regions are created by middle layer or hyperbox layer. Output layer comprises the number of nodes equals to the number of output classes [18].

In our proposed model, MFO generates optimal set of features to be supplied to classifier FMMNN for achieving classification accuracy with less error.

II. Review of Literature

Optimization algorithms are assorted into a number of candidate solutions and dependency criteria. Here number of candidate solutions are divided in two groups i.e. group based on Individual and group based on population. Simulated Annealing and Hill Climbing are some of the example of Individual based algorithm. Local optima is the basic problem with these algorithms. Brain Storm Optimization, Particle Swarm Optimization and Moth Flame Optimization (MFO) belongs to the category of population based and have the high capability to avoid problem of local optima [21]. Evolution, Physic and Swarm are three categories which comes under population based feature selection algorithms. With respect to the nature, evolutionary algorithm simulates its evolutionary process. One instance of the evolutionary computation is the Bio Geography Based Optimization (BBO) algorithm [6]. Gravitational Search Algorithm is inspired from physical phenomena in nature [7]. Social behavior of animal species like ant and birds, impacted swarm based algorithms. Some of its examples are Bees Algorithm [5], Grey Wolf Optimizer [21] and Moth Flame Optimization [21].

Two conflicting milestones i.e. Discovery and Exploitation should be balanced to give optimum result in optimization algorithm. Population based algorithms have the potential to balance discovery and exploitation in order to first find a reasonable approximation of the global optimum then boost its accuracy[2].

Optimization algorithms are sectioned into a filter based and wrapper based feature selection, based on its dependency criteria. In filter based perspective, data dependent criteria is used for getting possible solutions in the feature space whereas classification dependent criteria is used in wrapper based perspective for getting solutions in the feature space[8].

MFO is the one of the approach belongs to category of population and wrapper based feature selection. MFO algorithm finds very competitive outcomes compared with other well-known meta-heuristic algorithms such Particle Swarm Optimization and Biogeography based algorithm. MFO algorithm diversification is very high consequently it requires local optima to be avoided. Its Diversification and intensification balance is very efficient in seeking the right solution to address real issues [21].

A structure can be designed to establish the connectivity between optimization algorithm and the problem to get optimum solution [4]. Many researchers already worked with classifier like Neural Network, Fuzzy Logic, Hybrid Neurofuzzy model, support vector machine, principal component analysis, Artificial Immune System and Genetic Algorithm. Rüdiger W. Brause (2001) presented disease diagnosis model by the use of growing neural networks and rule based networks, he detailed neural network applicability to classify medical problems and showed its merits and demerits with regards to the medical background. He concluded that human diagnostic ability are worse than the neural network diagnostic systems [19] Lotfi Jadeh worked on fuzzy set theory and in 1974 presented Fuzzy logic and its application to approximate reasoning and stated that fuzzy logic is capable for making rule based intelligent system[12].During the year 1992, P. K. Simpson proposed two separate neural network training approaches: one for the problem of clustering and another for the problem of classification, Fuzzy min max neural networks were presented for their experiments [18]. Work on "Least square support vector machine with fuzzy weighing pre-processing" presented by E. Comak and achieved accuracy was 94.29% for the liver disorders problem [3]. It is noted that SVM does not accomplish the classification task in the non-linear case but in linear case SVM classifies the data. To support nonlinearity data needs to be converted in linear feature space by using different compatible kernels. SVM is a statistical learning theory and offline mechanism to develop medical decision support system. Lukka et al. suggested a 66.50 percent classification accuracy based on fuzzy robust PCA algorithms and similarity classifier for liver disorders [13]. Here, data has different point scales. So firstly there is need to normalize the data then it is to be applied to PCA. PCA is used for dimensionality reduction and it is not influenced by variables which are in high magnitude. Seral Ozens et al. experimented combination of genetic algorithm and artificial immune system for heart and liver disorders" [17]. The key downside here is that there is no classification bias for many AIS classifier. The pure distance criterion required to determine the degree of affinity. So AIS is associated with genetic algorithms in this approach to handle above mentioned problem. It is point to be noted that GA uses probabilistic rules to guide searching. Manjeevan seera et al. (2014) [15]", used "Hybrid Intelligent system for medical data classification (Fuzzy Min Max-CART-RF)". They experimented FMM, FMM-CART and FMM-CART-RF for getting classification accuracies from the labeled dataset of liver disorders and pima Indian diabetes. In the classification stage here, test samples fall between various groups in the overlapping/ and/or containment regions. The authors used the contraction strategy to solve this problem. In contraction approach, principle of minimal disturbance is used to avoid the problem of overlapped and/or containment regions. Here some amount of misclassification arises because the use of principle of minimal disturbance. Lukka (2011) [14], gave the model "Feature Selection using fuzzy entropy measures with similarity classifier" for evaluating classification accuracy for Pima Indian Diabetes. Here feature selection is applied on high dimensional medical dataset. Proposed algorithm is used to select the appropriate and optimal features but forgotten to increase the accuracy constraint of the classifier. The feature selection algorithm which supports both binary dataset and the multi class dataset sometimes produces high accuracy on the binary dataset but gives low accuracy when it is used in the multi class data set. Orkcu & Bal (2011) [16], used Binary Coded GA, BP and Real Coded GA for the disease dataset of Pima Indian diabetes. Highest 77.10 % accuracy was achieved in the consideration of Real/Actual Coded Genetic Algorithm. The actual Coded and Binary Coded Genetic Algorithm operated explicitly with character strings representing the parameter collection and not the parameters themselves. It also uses probabilistic rules not deterministic rules to guide their search.

Zhiying Xu et al. (2020) presented a soft computational model for diagnosing skin cancer. Here convolution neural network is optimized by the satin bowerbird optimization (SBO) algorithm hence overall classification accuracy enhanced when optimal features are classified by Support vector machine. As performance metric was concerned they achieved 95% accuracy [23]. Navid Razmjooy et al discussed implantation capabilities based model for the automatic skin cancer detection [26]. As per the study of authors Ana Carolina Borges Monteiro et al., Software Defined Radio embedded with Wireless Body Area Network system proved beneficial for doctors to keep an eye on patient. As this technology can measures level of blood pressure, temperature etc. This type of technological advancement proved very important in healthcare field [24]. Benefits of Health 4.0 cyber physical system (HCPS), elaborated by Ana Carolina Borges Monteiro et al. in healthcare system. The HCPS observes parameters of patient's condition using biosensors [25]. Study of related articles helped to motivate for the research in the medical diagnosis field. Advancement in medical diagnosis with optimized performance needed to equip with latest technology to maintain smart healthcare system.

As main findings are concerned, FMMNN suffers from problem hyperboxes overlapping and containment. For solving this, contraction approach is used. It is observed that some misclassification still exists here if too many features are handled while classification process [1]. MFO generate selected optimal features to be supplied to FMMNN. Hence overlapping and containment cases between hyper boxes in FMMNN reduced. Consequently chances of misclassification minimized. This is the main motivation of using combination of FMMNN and MFO in our model.

III. Moth Flame Optimization (MFO)

Moths navigate in a fixed angle targeting the moon. In night they are unable to distinguish between moon and artificial light or flame. For movement they follow transverse orientation navigation strategy also maintains a similar angle with flame and the moon [21]. Compared to the moon, flame remains highly closed to the moths so maintaining a similar angle triggers spiral navigation route for moths [10] hence convergence towards flames occurred.

A. MFO Algorithm

It is assumed that moths are the matrix of possible solutions and their layout in the space are the problem variables. The best obtained layout (optimal position) of moths is kept in matrix of flames [21]. Following paragraphs shows the overview of steps of MFO algorithm.

Step 1: Required main Parameters: count of moths (search agents), count of flames, count of variables, count of maximum iterations, lower bound (lb (i)) and upper bound (ub (i)) of variables.

Step 2: Representation of Population: MFO algorithm can be modelled by the position of moths and flames. The MP matrix represents the position matrix of moths.

$$\text{MP}=\begin{pmatrix} \text{MP}_{1,1}\ \text{MP}_{1,2}\ldots\text{MP}_{1,d} \\ \text{MP}_{2,1}\ \text{MP}_{2,2}\ldots\text{MP}_{2,d} \\ \text{MP}_{n,1}\ \text{MP}_{n,2}\ldots\text{MP}_{n,d} \end{pmatrix} \tag{1}$$

here n is the count of moths and d is the count of variables (dimensions).$\text{MP}_{i,j}$ Shows the value of jth position variable of the ith moth, is decided by Eq. (2)

$$\text{MP}_{i,j} = (\text{ub}(i)\text{-lb}(i) * \text{rand}(\ ) + \text{lb}(i)) \tag{2}$$

here rand ( ) is the random number generator in the interval [0, 1] having random distribution. FP represents the position matrix of flames as follows:

$$\text{FP}=\begin{pmatrix} \text{FP}_{1,1}\ \text{FP}_{1,2}\ldots\text{FP}_{1,d} \\ \text{FP}_{2,1}\ \text{FP}_{2,2}\ldots\text{FP}_{2,d} \\ \text{FP}_{n,1}\ \text{FP}_{n,2}\ldots\text{FP}_{n,d} \end{pmatrix} \tag{3}$$

$\text{FP}_{i,j}$ Shows the value of jth position variable of the ith flame.

In MFO, moths are the search agents that move around the search space and flames are the best position of moths.

Step 3: Persistence of fitness function: Fitness function evaluates each month. FOM matrix stores fitness value of moths processed by fitness function (see section 3.2). Similarly the matrix FOF is used to store the fitness values of flames.

$$\text{FOM} = \begin{pmatrix} \text{fom}_1 \\ \text{fom}_2 \\ \text{fom}_n \end{pmatrix} \tag{4}$$

$$\text{FOF} = \begin{pmatrix} \text{fof}_1 \\ \text{fof}_2 \\ \text{fof}_n \end{pmatrix} \tag{5}$$

In MFO, Flames are treated as flags dropped by moths while searching the search space. In the event of seeking a better solution, each moth looks for a flame and updates its position .Flame matrix updates their fitness values as per the fitness values of moths.

Step 4: General texture of MFO and iteration process: - The general texture of MFO algorithm is as follows

$$MFO = (K,R,T) \tag{6}$$

where, K is the function that generate a random population of moths and corresponding fitness, R is the main function that makes the moths move around search space, T is a termination function. Function R uses logarithmic spiral, it is main update mechanism for moths to simulate orientation behaviour of moth. Eq. (7) update location value of moth with respect to the flame.

$$MP_i = S(MP_i, FP_j) = D_i . e^{bt} . \cos(2\Pi t) + F_j \tag{7}$$

where, the spiral function is S, MP_i is the ith moth, and FP_j is the jth flame, D_i defines the distance between the ith moth and the jth flame for defining the shape of logarithmic spiral. b is a constant and t is a random number in the range [-1, 1]. D_i is measured using following Eq. (8)

$$\text{D}_i = |\text{FP}_j - \text{MP}_i| \tag{8}$$

Spiral function $S(MP_i, FP_j)$decides how moths update their position around flame. Eq. (7) allows a moth to fly around a flame only. Hence the exploration/discovery and exploitation of the search area are therefore ensured. In order to emphasize exploitation further it is believed that t is a random number in range [r, 1] where r is the constant of convergence. Its values linearly decreased from -1 to -2 over the course of iteration. But exploitation of the best solutions may degrade due to update of moth's position with regard to n various positions in the search space. Because of this reason, MFO can fall in local minima. To solve this problem Eq. (9) is used to reduce the quantity of flames over the course of iterations. Hence exploitation degradation achieved and best promising solution obtained.

$$\text{Flame count = round } (NF - c \times \frac{NF-1}{T}) \tag{9}$$

where, NF is the maximum count of flames, c is the current iteration, and T is the maximum count of iterations.

Step 5: Obtaining optimal solution: - General steps of R function are as follows:

```
Update flame count using Eq. (9)
FOM = FitnessFunction (MP);
 If iteration == 1
        FP = sort (MP);
        FOF = sort (FOM);
else
        FP = sort(MPt-1, MPt);
        FOF = sort(MPt-1, MPt);
end
for i = 1: n
        for  j = 1: d
            Update  r and t
            Calculate D using Eq. (8) with
            respect to the corresponding moth.
            Update MP (i,j) using Eq. (7) with
            respect to the corresponding moth.
        end
end
```

R function is executed until T function i.e. T: MP →{true, false} returns true. After termination of the R function the best moth is returned as the best obtained approximation of the optimum.

B. Objective Function (Fitness Function) used in MFO

The aim of the optimization has to improve the classification accuracy of the classifier by selecting minimum set of the attributes. This improves the accuracy of the system with improved speed of classification which is achieved by the reduction of attributes using MFO method. Let $[X_1 \ldots\ldots\ldots.. X_n]$ are n attributes of data sample (disease data sample). If k attributes selected out of n then optimization function have k variables $[X_1 \ldots\ldots\ldots.. X_k]$ Range of each variable is as follows.

$1 <= g_k <= n$ where constraint is $g_i \neq g_j$ for all value of i and j.

In this way we select the reduced attribute set as $A_R = [X_{g_1} \text{.........} X_{g_k}]$ where A_R is set of reduced attributes. Following equation represent the main motive of the objective function. It is formed as to reduce classification error by FMMNN classifier for A_R.

$$\text{minimized classification error} = \text{FMMNN}(A_R, T) \qquad (10)$$

where T is the target output.

FMMNN classifier consist three layers. In the first (Input) layer, positions of moths are supplied as the parameters. In the middle layer, formations of hyperboxes (logically separated regions as per its class) take place. Its output layer gives the actual classified output. Then classification error is calculated based on the comparison between target output and actual output. Section 4 gives the more details about FMMNN. In MFO, values of optimal (efficient and reduced) set of attributes (where minimum classification error is obtained) considered as fitness value of moths.

IV. Fuzzy Min Max Neural Network (FMMNN)

F_A, F_B and F_c are the three consecutive layers of FMMNN Fig. 1, where F_A represents input layer, F_B acts as hidden or hyper box layer and F_c is the representative output layer which contains class nodes. In hyper box layer, a hyperbox is regarded as a fuzzy set. The minimum and maximum points of the hyper boxes are denoted by matrices V and W respectively, matrix U gives relation between hyperbox layer and output layer.

A. Transition from Input Layer to Hyperbox Layer

FMMNN uses membership function as given in eq. (11) to calculate input features belongingness for the hyper boxes

$$b_j = (a_j, V_j, W_j) = \frac{1}{n}\sum_{i=1}^{n}(1 - f(a_{h_i} - W_{j_i}, \gamma) - f(V_{j_i} - X_{h_i}, \gamma)) \quad (11)$$

$f_{x,y}=1$ if $x*y>1$

$f_{x,y}=x*y$ if $0<=x*y<=1$

$f_{x,y}=0$ if $x*y<0$

here a_h = Input features, n = dimension of feature space, b_j = j^{th} hyper box, V_j = set of min points, W_j = set of max points, f(x, y) is a two parameter function where γ is sensitivity(fuzziness control) parameter. The connections between input layer and hyperbox layer maintained by v and w matrix.

In hyper box layer, if training performed 1st time creates a hyper box using $V_j = W_j = A_h$ else expand existing hyperbox b_j using new training data if the criteria given in Eq. (12) satisfied:

$$\theta \geq \sum_{i=1}^{n}(\max(w_{j_i}, a_{h_i}) - \min(v_{j_i}, a_{h_i})) \qquad (12)$$

here θ = Expansion coefficient

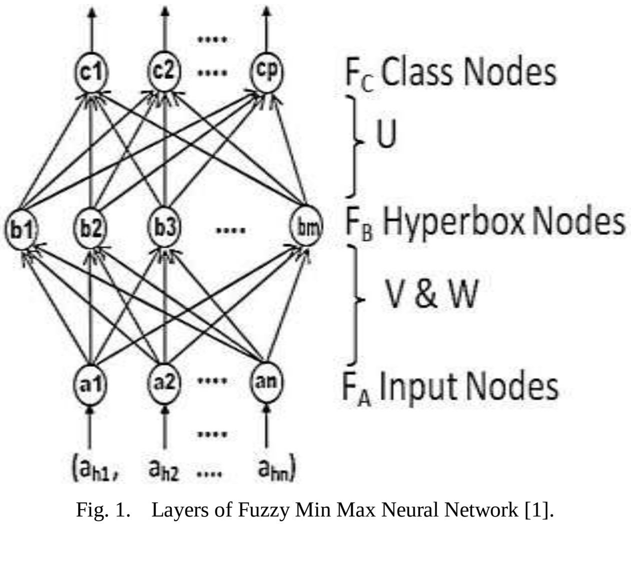


Fig. 1. Layers of Fuzzy Min Max Neural Network [1].

When a training data is given to the network then it try to accommodate in one of the existing hyperbox of that class provided that the hyperbox size is not exceeding the specified maximum limit. To know the overlapping or containment between hyperboxes following tests are to be done. Assume that hyperbox b_j and b_k are compared with each other then

a) Isolation test: If $\{W_{j_i} < V_{k_i}\}$ or $\{W_{k_i} < V_{j_i}\}$ true for any value of i, it shows that two hyperboxes (b_k,b_j) are isolated.

b) Containment test: If $\{V_{j_i} < V_{k_i} < W_{k_i} < W_{j_i}\}$ or $\{V_{k_i} < V_{j_i} < W_{j_i} < W_{k_i}\}$ true then b_j is contained by b_k or b_k is contained by b_j.

c) Overlap test: if test (a) and test (b) are not satisfied then overlapping between b_k and b_j is obvious.

We used contraction approach to remove hyperboxes overlapping and containment problem. In this, if the hyperboxes from different classes overlap, a hyperbox contraction process is initiated to eliminate the overlaping regions. Note that overlapping regions caused by hyperboxes from the same class are allowed [1].

B. Transition from Hyperbox Layer to Output Layer

Connection between hyperbox node and class nodes in is represented by u matrix.

$u_{ij} = 1 \quad if \{b_j \varepsilon c_i\}$ otherwise $u_{ij} = 0$

By knowing the hyperboxes belongingness to the class, network easily classify the linearly non separable data. c_i decides the membership of i^{th} class and it is given by

$$c_i = \max_{j=1}^{m}\left(b_j u_{ij}\right) \tag{13}$$

V. Methodology of MFO-FMMNN

Our proposed hybrid model consist two parts. First part handled by Moth Flame Optimization phase (MFO) while the second part took over by the classification phase. Methodology structure of hybrid model is shown in Fig. 2. Methodology steps are described below:

1) Input:All features of medical (disease) data samples are supplied to MFO-FMMNN.

2) Settings of parameters: Following parameters need to be set in this step. (i) count of moths(search agents) (ii) count of flames (iii) count of variables (iv) count of maximum iteration (v) vector of Lower bound vector (vi) vector of upper bound. (3) Setting of population of moths and count of flames:

3) Origin of moth and flame count: Set the first random population of moth, obtained from the Eq. (2). Count of flames is calculated using Eq. (9).

4) Inception of optimization:

a) MFO started with iteration value 1.here iteration is the count of loops used to test algorithm.

b) The fitness values of all the moths calculated using objective function (see Section IIIB, the values which produced minimum error in its extent when medical features processed by FMMNN as fitness f unction is concern)

c) Sorting of the first population of moths takes place based on their fitness values.

d) The population with the best moths fitness values will be selected as the flames hence the fitness values of flames and moths equals

e) Updating of moths position takes place with respect to its corresponding flame, calculated according to Eq. (7).

5) Culmination of optimization:

a) When the first iteration terminates we get the best set of moth and flames and the best fittest values of moths and flames.

b) Start to generate the offspring generation of moths (other subset of features) then update the position and fitness values of moths and flames at each subsequent iterations according to (Eq. (1) to (5) and (7))

c) In each iteration the sequence of flames are Modified based on the best solutions, then moths update their position with respect to the updated flame.

d) When iteration value reaches its maximum limit, we get the new best set of moths and flames and also their corresponding fitness value. Now optimization process stops and the best moths (optimal set of features) obtained.

6) Classification: Using optimal set of features, FMMNN classifier calculates classification accuracy.

a) In MFO phase, FMMNN runs iteratively with regards to obtain fitness values of moths as per iterations.

b) In classification phase, FMMNN runs to obtain classification accuracy.

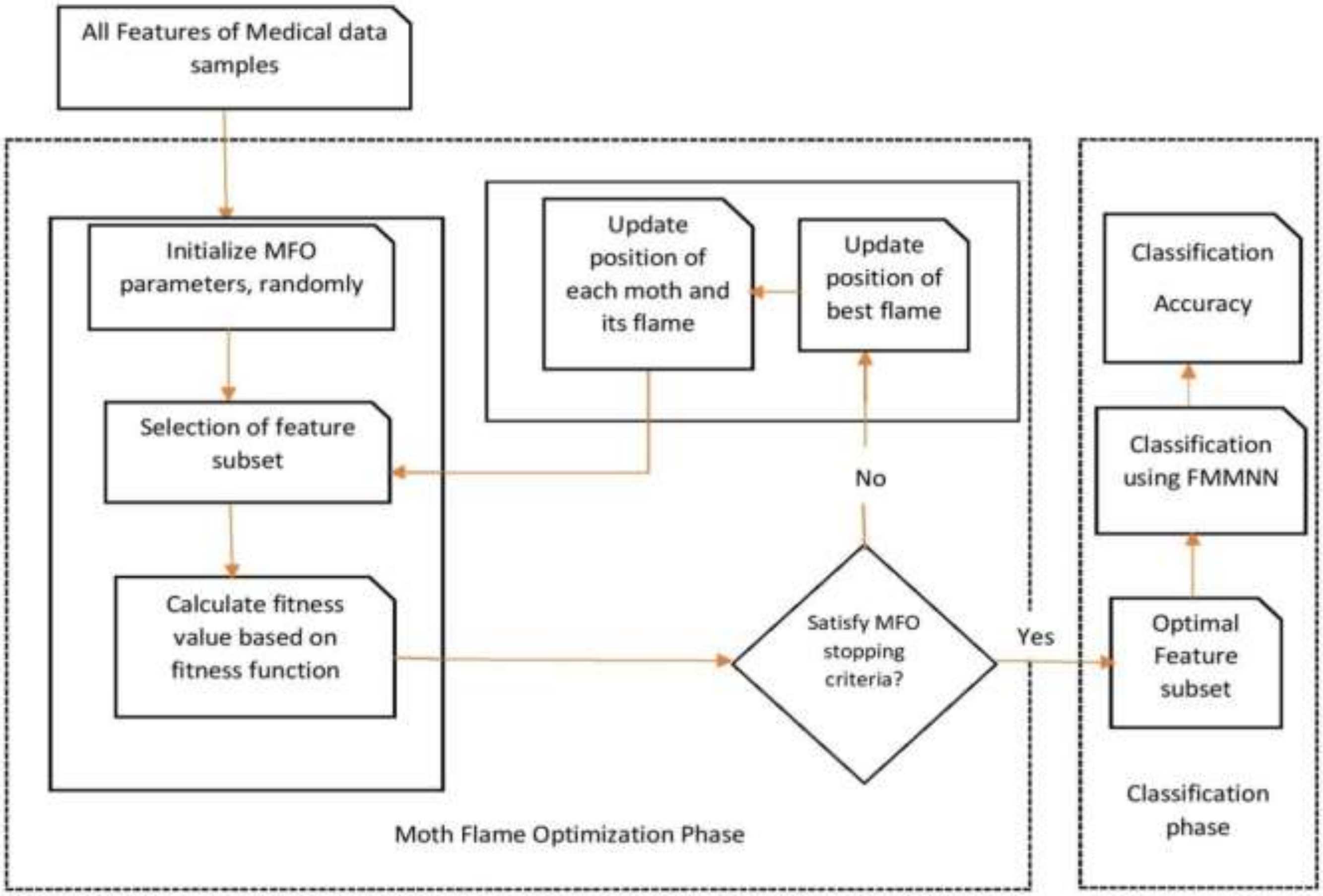

All Features of Medical data samples
Initialize MFO parameters, randomly
Selection of feature subset
Calculate fitness value based on fitness function
Update position of each moth and its flame
Update position of best flame
No
Satisfy MFO stopping criteria?
Yes
Moth Flame Optimization Phase
Classification Accuracy
Classification using FMMNN
Optimal Feature subset
Classification phase

Fig. 2. Methodology Structure of MFO-FMMNN.

VI. Experiments and Result Analysis

Our experiments and results are based on data samples of Liver Disorders and Pima Indian Diabetes. Following paragraphs shows the deliberation.

A. Experiment on Liver Disorders Dataset

Publicly available data sets of Liver Disorders taken from the University of California at Irvine (UCI), machine learning data repository (This repository consist genuine database and can be considered for research [11]). It is used for evaluation of medical decision support system. In the dataset following feature values of the patient are considered i.e. Age, Gender , Bilirubin Total, Bilirubin Direct, Alkaline Phosphotase, Alamine Aminotransferase, Aspartate Aminotransferase, Protiens Total, Albumin, Albumin and Globulin Ratio(Lichman et al. , 2013). A series of systematic evaluations was conducted for 1 to 10 features for 50 Maximum count of iterations. Fig. 3 shows result of percentage of classification accuracy evaluated by MFO-FMMNN. When 5 features are considered, achieved accuracy 97.74 %. This result is compared with the result calculated by FMMNN classifier then it found that MFO-FMMNN proves better than FMMNN as per accuracy result shown in Table I.

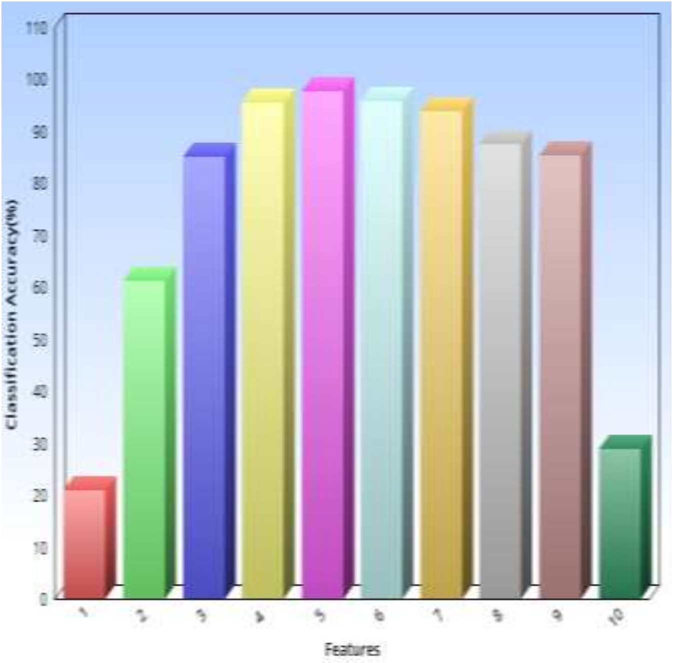


Fig. 3. Features vs. Classification Accuracy Calculated by MFO-FMMNN for Liver Disorders Dataset.

TABLE I. COMPARISON OF CLASSIFICATION ACCURACY BETWEEN FMMNN AND MFO-FMMNN FOR LIVER DISORDERS DATASET

SN	Model/methods used	Accuracy achieved
1	FMMNN	68.80
2	MFO+FMMNN	97.74

B. *Experiment on Pima Indian Diabetes Dataset*

This dataset is taken from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to classify whether or not a patient suffered from diabetes [22]. Patients data considered here are of females at least 21 years old of Pima Indian heritage. Total eight features are considered in each data sample i.e. patient age, plasma glucose concentration a 2 h in an oral glucose tolerance test, blood pressure (diastolic), triceps skin fold thickness, 2-h serum insulin, index of body mass, diabetes pedigree function of diabetes, times of number of pregnant. When PID dataset is processed by MFO-FMMNN for 70 Maximum numbers of iterations then calculated classification accuracy is 86.95 % (see Fig. 4). Our output is compared with the classification result calculated by previous author's proposed work. Results are summarized in Table II, which represents the comparison of classification accuracy between FMMNN and MFO-FMMNN for Pima Indian diabetes dataset. Accuracy achieved by MFO-FMMNN super shaded FMNNN.

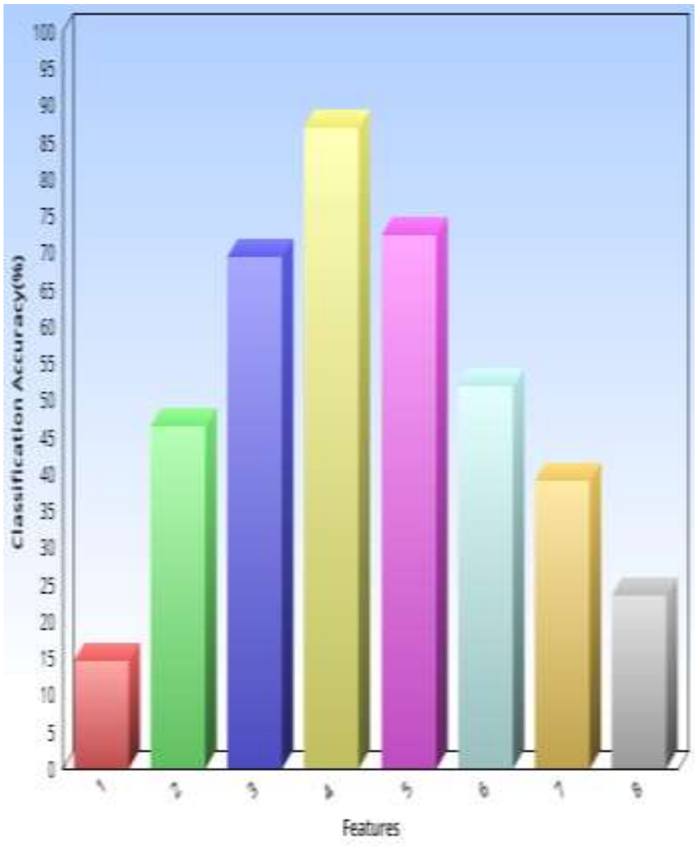


Fig. 4. Features vs. Classification Accuracy (%) Calculated by MFO-FMMNN for Pima Indian Diabetes Data Sample.

TABLE II. COMPARISON OF CLASSIFICATION ACCURACY BETWEEN FMMNN AND MFO-FMMNN FOR PIMA INDIAN DIABETES DATASET

SN	Model/methods used	Accuracy achieved
1	Fuzzy Min Max Neural Network	70.40
2	MFO+FMMNN	86.95

VII. CONCLUSION

MFO-FMMNN can solve difficult problems in the real world with restricted search spaces. MFO balances discovery and exploitation of search spaces hence able to avoid local minima problem. This is the main reason behind getting optimal features set. FMMNN used in this paper fulfilled two purposes i.e. calculating fitness value of features (Moths) in MFO phase and classify the data in single pass classification phase. It has the property of incremental learning such that newly introduced data can be classified easily. Presented hybrid model is able to calculate optimal accuracy by considering different sets of features of supplied dataset hence best accuracy from best set of features guaranteed. The limitation of this model is that MFO method is not capable to generate solutions for problems having more than one objective. Decision trees can be used to provide explanatory rules of working of this hybrid model. In future this model can also be implemented and tested for image basis medical datasets if wavelet transforms accompanied with it.

REFERENCES

[1] A.V. Nandekar and P.K. Biswas, "A Fuzzy Min-Max Neural Network Classifier With compensatory Neuron Architecture," IEEE TRANSACTIONS on neural networks, vol. 18, no. 1, January 2007.

[2] A.E. Eiben, C. Schippers, On evolutionary exploration and exploitation, Fundam. Inform. 35(1998) 35–50.

[3] E. Comak et al (2007). A new medical decision making system: least square support vector machine (LSSVM) with fuzzy weighing preprocessing, Expert system with applications 32(2), 409-414.

[4] D.H. Wolpert, W.G. Macready, No free lunch theorems for optimization, IEEE Trans. Evol.Comput. 1 (1997) 67–82.

[5] D. Pham, A. Ghanbarzadeh, E. Koc, S. Otri, S. Rahim, M. Zaidi,The bees algorithm-a novel tool for complex optimisation problems, IPROMS 2006, 2006, pp. 454–459.

[6] D. Simon, Biogeography-based optimization, IEEE Trans. Evol. Comput. 12 (2008) 702–713. (2014) 46–61.

[7] E. Rashedi, H. Nezamabadi-Pour, S. Saryazdi, GSA: a gravitational search algorithm, Inf. Sci. 179 (2009) 2232–2248.

[8] I. Guyon, A. Elisseeff, "An introduction to variable and attribute selection", Machine Learning Research,Vol.3, pp. 1157-1182, 2003.

[9] Hossam M. Zawba et. al, "Feature Selection approach based on moth flame optimization algorithm" 978-15090-0623-6/16/$31.00©2016 IEEE

[10] K.D. Frank, C. Rich, T. Longcore, "Effects of artificial night lighting on moths", Journal of Ecological consequences of artificial night lighting, Hindawi Publishing Corporation, Vol. 2015, No. 1,pp. 305-344, 2006.

[11] Lichman, M. (2013). UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science

[12] Lotfi jadeh(1974). "Fuzzy logic and its application to approximate reasoning". In: Information Processing 74, Proc. IFIP Congr. 1974 (3), pp. 591–594.

[13] Luukka at al.(2009). Classification Based on Fuzzy Robust PCA algorithms and similarity Classifier. Expert Systems with Applications, Elsevier, 5 March 2008.

[14] Luuka P.(2011) Feature selection using fuzzy entropy measure with similarity classifier. Expert system with application.38 (4), 4600-4607.

[15] Manjeevan Seera ,CheePeng Lim. “A hybrid intelligent system for medical data classification”, ELSEVIER, Expert System with application 41(2014) 2239-2249.

[16] Orkcu H. H. & Bal H.(2011) Comparing performance of backpropagation and genetic algorithms in data classification. Expert system with applications 38(4),3703-3709.

[17] Ozens S & Gunes (2009). Attribute weighing via genetic algorithm for attribute weighted artificial immune system (AWAIS) and its application to heart disease and liver disorders problems. Expert system with applications, 36(1), 386,392.

[18] P. K. Simpson, “Fuzzy min-max neural networks –Part 1: Classification”, IEEE Trans. Neural Networks, Vol.3, No.5, pp. 776-786, 1992.

[19] R. W. Brause, “Medical Analysis and Diagnosis by Neural Networks”, J.W. Goethe - University,Computer Science Dept., Frankfurt a. M.,Germany,2001.

[20] S. Mirjalili,” Moth - Flame Optimization Algorithm: A Novel Nature inspired Heuristic Paradigm”, Knowledge-Based Systems, Vol. 89, pp.228-249, 2015.

[21] S. Mirjalili, S.M. Mirjalili, A. Lewis, Grey wolf optimizer, Adv. Eng. Softw. 69.

[22] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press.

[23] Zhiying Xu*, Fatima Rashid Sheykhahmad, Noradin Ghadimi*, Navid Razmjooy (2020). Computer-aided diagnosis of skin cancer based on soft computing techniques. Open Medicine, 15(1), 860- 871.

[24] Ana Carolina Borges Monteiro , Vania V. Estrela , Abdeldjalil Khelassi , Yuzo Iano, Navid Razmjooy , Diego T. M. Rocha , Delcimar Martins (2019). Why software-defined radio (SDR) matters in healthcare? Med. Technol. J. 3 (3), 421–429.

[25] Ana Carolina Borges Monteiro , Vania V. Estrela , Abdeldjalil Khelassi , Yuzo Iano, Navid Razmjooy. (2019) Health 4.0: Applications management technologies and review type of article. Med. Technol. J., 2(2), 262-276

[26] Navin Razmjooy, Mohsen Ashourian, Maryam Karimifard, Vania V. Estrela, Hermes J. Loschi, Douglas do Nascimento, Renaldo P. Franca, Mikhail Vishnevski(2020), Computer-added Diagnosis of skin cancer: a review.Current medical Imaging, 16(7), 781-793.

A Preliminary Intergenerational Photo Conversation Support System based on Fine-tuning VGG16 Model

Lei JIANG[1], Panote Siriaraya[2], Noriaki Kuwahara[4]
Graduate School of Science and Technology
Kyoto Institute of Technology
Kyoto, Japan

Dongeun Choi[3]
Faculty of Informatics
The University of Fukuchiyama
Kyoto, Japan

***Abstract*—China has the largest number of elderly people in the world, young volunteers have become the main force in caring for the elderly. It is urgent to establish a photo conversation support system to build a bridge of communication between young volunteers and the elderly. Previous research generally used perceptual analysis or machine learning methods to find photos suitable for intergenerational conversation, this paper uses deep learning models to further learn the potential features of two datasets suitable for and not suitable for intergenerational conversations. However, the original datasets are too small, it was first proposed to use TF-IDF in conversation recording and data augmentation technology in images to expand the datasets. Then on the basis of the VGG16 model combined with transfer learning and fine-tuning technologies, five models were designed. The accuracy of the best model on the validation set and test set reached 96% and 94.5%. In particular, the recall rate of the not suitable dataset reached 100%, all not suitable photos were identified. At the same time, the recall rate of other datasets reached 71% for not suitable photos. It shows that the system is also applicable to other datasets and can effectively eliminate photos that are not suitable for intergenerational conversations.**

Keywords—Intergenerational photo conversation support system; TF-IDF; VGG16; transfer learning; fine-tune

I. Introduction

A. Changes in Care Models and Cognitive Impairment

China has the largest aging population in the world. We urgently need to address the problem of improving quality of life for these senior citizens. , Currently, there are three modes of family care in China: institutional care and community home care [1]. As the proportion of 4-2-1 structured households (four seniors, a couple and one child) increases year by year, couples have little energy to allocate to the elderly, and it is difficult to meet the emotional needs of the elderly on a purely material basis [1]. Although institutional old-age care promotes the “integration of medical and nursing care” [2], and the companionship of other elderly people can reduce loneliness, China's experience in changing from an adult society to an aging society has been relatively fast, and the number of places is insufficient. On average, there are only 21.5 places per 1,000 senior citizens [3]. Community home care is provided by the community based around the family being the core provider of related care services for the elderly living at home. It is an effective long-term care method that is popular among the elderly. However, there are also some problems. There is a shortage of professional personnel for care giving, and the majority of the workforce is made up of volunteers. As of 2012, 33.92 million people had registered as youth volunteers[4]. However they only account for 2.54% of the total population and few people regularly participate in voluntary activities for a significant amount of time, mostly due to lack of communication and emotional exchanges[4]. If the volunteers lack experience in caring for and communicating with the elderly, it can become difficult for the two generations to empathize and find common ground. At the same time, the increase of non-traditional family structures such as dink families (double income and no kids), lost families, and empty nest families means that for 47.53% of the elderly in China in 2016 are in these non-traditional family structures[5]. 60% of the elderly in non-traditional family structures have psychological problems related to cognitive impairment[6]. Strengthening daily communication to ensure good social relations and minimizing psychological distress may help delay or prevent cognitive impairment[7].Therefore, it is extremely important to find a way to help young people communicate with the elderly and help them to establish emotional connections.

B. Research Purpose

This paper aims to establish a preliminary intergenerational conversation support system which is able to help the elderly and young volunteers to quickly and effectively screen their favorite photos to inspire conversation.

C. Research Materials

In the early stage, our research team has done two generations of conversation experiments supported by photos, and grouped all the photos into groups that are suitable for the conversation of the elderly and not suitable for the elderly. The research in this paper is based on the results in [8]. Through the deep learning model to learn the characteristics of each group of photos. So that it can effectively eliminate photos that are not suitable for two-generation conversations and obtain suitable photo.

The first group of photos are the third cluster of [8] which are suitable for the elderly to talk about. It is classified as the good category that inspired many conversation topics and made conversation flow freely between the two generations. These contents are: “cakes”, “lotus flowers”, “shared bicycles”, “red dates and white fungus soup”, “pandas”, “foot therapy”, “movie theaters”, “Water Cube (Beijing National Aquatics

Center)", and "outings". The second group of photos are the first cluster of the paper which are not suitable for the two generations communication. It is classified as the bad category, as the content of the conversation is limited and both parties are stressed. These contents are: "Shengjianbao", "Zhang Yimou and Gong Li", "Qinhuai River", "mother teaches children to fold clothes", "Jackie Chan", "toothpaste", "rape flowers", and "gourds".

Although the experimental photos are divided into two groups, it is difficult to distinguish the category of each group of photos by traditional statistical methods. For example, "raw fried buns", "cakes", "red dates and white fungus soup" all belong to the food category, but the elderly and young people have different attitudes towards them. "Rapeseed" and "lotus" both belong to plants, but the attitudes towards them are also different. Only after conducting enough dialogue experiments can we understand the relationship between the groups and the photos within the groups. This process is very time-consuming and labor intensive. In order to overcome this difficulty, we chose to use deep learning models to help us learn the potential relationships between and within groups, provide us with guidance to choose photos for the next set of conversations, and to initially establish a photo conversation support system. The main contributions of this paper present the following:

- In order to solve the problem of insufficient data set for the deep learning models , not only the commonly image processing techniques used in photos such as rotation and stretching, but also the TF-IDF technology used to expand the content of the photos to achieve the expansion of the data set purpose.
- On the basis of the VGG16 model, five models combined with transfer learning and fine-tuning technologies were designed. Compared with the CNN model trained from scratch, it greatly improves the accuracy and reduces the amount of calculation and training time.
- Data sets from other regions are used to explore the applicability of the model of this system.

Rest of the paper is organized as follows. Section II introduces an overview of the existing research. Section III presents data acquisition, date augmentation, model architectures, model evaluation and the process of establishing the intergenerational photo conversation support system. Results obtained the model which is able effectively eliminate photos that are not suitable for two-generation conversations in Section IV. Section V presents the discussion of this paper. The conclusion and perspectives of this paper are provided in Section VI.

II. Related Work Reviewer A Q3

Astell et al. [9-12] developed CIRCA, a multimedia system displayed on a touch screen for dementia patients. The project started in 2001 and now includes photos, videos, music, graphics and text. CIRCA is designed to be used by dementia patients and caregivers, and it provides cognitive support for people with dementia covering communication, entertainment and creativity. Alm et al. [13] carried out a long-term trial and evaluation of CIRCA, expanded the multimedia content and introduced a randomization function, and found that this was able to provide enough interesting content to stimulate dementia patients to recall memories that were previously unheard. CIRCA was developed in Dundee to explore the possibility of using the system in other regions, and Purves et al. [14] selected relevant content for seniors for a British Columbia version of CIRCA. The pilot tested CIRCA-BC on 3 participants with dementia and a conversation partner. By analyzing their interactions, they found that CIRCA's content can be adjusted for use in different regions based on similarities and differences in the participants' shared social history. Fels et al. [15] proposed using everyday photos to inspire dementia patients to share stories about their lives and life experiences in order to establish emotional connections. He found that regular daily conversations allow participants and listeners to have pleasant shared interactions. Miyuki Iwamoto et al. [16] researched a dialogue support system hoping to reduce the stress of young volunteers in the system when talking to the elderly through the shared content of photos and videos. The results of this were that the video content caused subjects to feel less stressed, but that the photos allowed conversation between older people and young volunteers to last for longer periods of time. Relatively little research has been carried out in China. Zhou et al. [17] took 40 photos and conducted 160 conversation experiments. He discovered that photos of "food", "event", "school" and "commodity" were suitable for conversation between two generations of Chinese. Zhou in another paper[8] used the PCA method and HCA to analyze and classify 40 photos, hoping to find the best photo cluster suitable for establishing an intergenerational photo dialogue support system. Although this research did not achieve a perfect result, it provided guidance for photo classification projects in the future.

From previous related researches, three main points can be concluded. The one is that photos make the talker more focused and more relaxed and can stimulate participants to share stories of personal life and life experience. so we used photos as the material for the intergenerational conversation is feasible. The other is that conversation system can be applied to different regions, if the materials are carefully selected. Therefore, our research in three different cities were conducted to explore the possibility of applying the intergenerational photo conversation support system in different regions. The last, previous researches generally used perceptual analysis or machine learning methods to find out the photos suitable for conversation and the deep learning models have never been used.

III. Research Methodology

A. Data Acquisition

Each group contains very few photos and as such, we cannot use deep learning models. We can instead expand the semantics of photos through natural language technology. Specifically, each photo contains four groups of dialogues, and we perform word segmentation and keyword extraction on the dialogues. We are able to use the keywords as candidate words to design a questionnaire, in order to first determine the topics of the candidate words that the two generations are and are not interested in, and then to select photos according to each of these topics. Finally, two types of photos are counted, and the photos in each category are randomly divided into a training set and a verification set at a 2: 1 ratio.

1) Keyword extraction: In order to dig deep into the topics that old and young people like and dislike, we analyzed their conversation content for each photo through TF-IDF and extracted the keywords. TF-IDF (term frequency-inverse document frequency) is a commonly used weighting technique for information retrieval and data mining [18]. It uses numerical statistics to reflect the importance of a word relative to a document in the corpus[19]. The first step is to calculate the TF:

$$\mathrm{TF} = \frac{\mathrm{count(t)}}{\mathrm{count}(\mathrm{d_i})} \qquad (1)$$

In the equation(1), count (t) represents the number of words contained in the document; count (d_i) represents the total number of words in the document. As can be seen above, if a word appears frequently in a document the value of TF is high.

The second step is to calculate the IDF:

$$\mathrm{IDF} = \frac{\mathrm{num(corpus)}}{\mathrm{num(t)+1}} \qquad (2)$$

In the equation(2), num (corpus) represents the total number of documents in the corpus; num (t) represents the number of documents in the corpus containing the particular words. As can be seen above, if a word rarely appears in other documents in the corpus, the value of IDF is low.

The third step is to calculate the TF-IDF:

$$\mathrm{TF—IDF} = \mathrm{TF(t)} \times \mathrm{IDF(t)} \qquad (3)$$

In the equation(3), the higher the TF-IDF of a word, the more it is considered to be important and representative for the documents. The TF-IDF algorithm can be implemented in any programming language of your choice. Because the Chinese document must be preprocessed using word segmentation and stop word filtering, before the TF-IDF calculation, this article uses a module named jieba with python for convenience. We calculated the TF-IDF value of the keywords of each photo in descending order, and selected the top five.

2) Data selection by questionnaire: Questionnaires were designed to confirm effective keywords and further explore the opinions of older people towards the photos. Table I shows the questionnaire participants. A total of two questionnaires were designed. Questionnaire one was used to let elderly people select keywords with positive impressions from fifty keywords that were extracted from the good category and then select positive pictures related to each keyword selected. Finally, the younger participants choose their positive pictures from the results of the elderly group. Questionnaire one includes two parts. The first ,for the elderly:

1. Do you have a positive impression of this keyword?

– Yes, I do. Please see question 2.

– No, I don’t. Please see the next keyword.

– I don't know. Please see the next keyword.

2. Do you like this photo related to the keyword?

– Yes, I do. (Keep this picture)

– No, I don't. (Delete this picture)

– I don't know. (Delete this picture)

The another for the young: Do you have a positive impression of this photo (Photos selected by the elderly group)?

– Yes, I do. (Keep this picture)

– No, I don't. (Delete this picture)

– I don't know. (Keep this picture)

Questionnaire two is only for seniors. The elderly person selects keywords with negative impressions from forty keywords extracted from the bad category and then selects the negative pictures related to each keyword selected. The content of the questionnaire two for the elderly:

1. Do you have a negative impression of this keyword?

– Yes, I do. Please see question two.

– No, I don't .Please see the next keyword.

– I don't know. Please see the next keyword.

2. Do you dislike this photo related to the keyword?

– Yes. I have. (Keep this picture)

– No, I don't.(Delete this picture)

– No, I don't know(Keep this picture)

In this way, there are no photos in the photo collection that the elderly do not like, and all of the photos are liked by both the elderly and young people.

B. Data Augmentation

All photos collected were divided via the questionnaires into a training set and a validation set in a 2: 1 ratio of the two categories, good and bad. We then augmented the training set using the keras preprocessing module ImageDataGenerator in python. There are many options available in ImageDataGenerator, but we chose just four of them and set Rescale = 1./255, Shear_range = 0.2, Zoom_range =0.2, Horizontal_filp = True. By using this tool our training sets were amplified tenfold. Rescale is a value by which we can multiply the data before any other processing[20]. Our original images consist of RGB coefficients in the range 0-255, so we target values between to 0 and 1 to instead by scaling with a factor of 1./255. Shear_range is for randomly applying shearing transformations. Zoom_range is for randomly zooming into pictures. Horizontal_filp is for randomly flipping half of the images horizontally relevant when there are no assumptions of horizontal asymmetry.

TABLE I. THE PARTICIPANTS OF QUESTIONNAIRE

No.	Questionnaire Location	Participant	Dementia
1	Shanghai Qibao Vanke City Garden Community	2 elderly females and 3 elderly males	No
2	Shanghai Qibao Vanke City Garden Community	2 young females and 1 young male	No

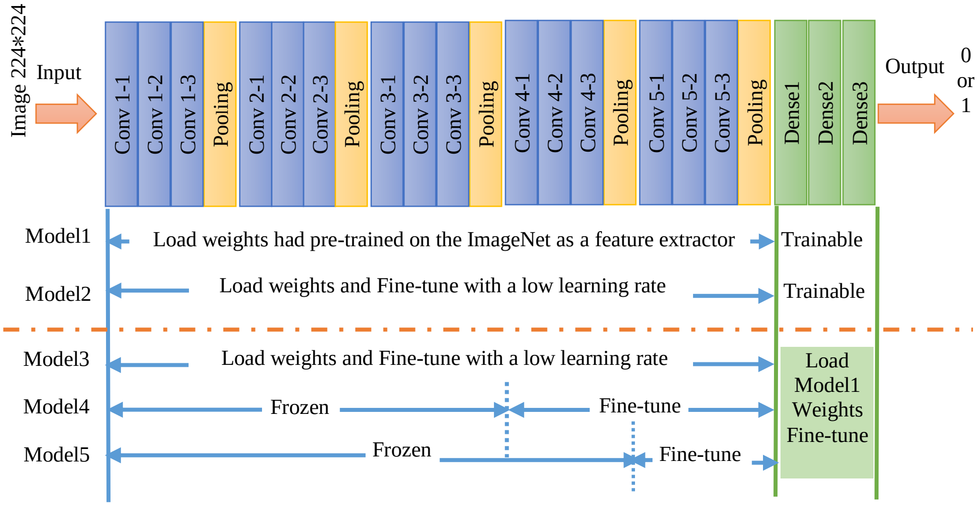


Fig. 1. VGG16 Model Architecture and Designed Models Architecture.

C. Model Design and Architecture

Although photos have been augmented on the content and number of photos, the data set is still too small for a deep learning model. For small amounts of data, data augmentation, transfer learning, fine-tuning, or a combination of several methods is generally used as in [21-24]. It used rotation, skewing and elastic distortion augmentation methods for images and then used a pre-trained CNN model as feature extractor and SVM as a category classifier. This technique has been applied in many fields and has shown better accuracy than traditional convolutional networks [25-27]. In this paper after we augmented the training set, we combined transfer learning and fine-tuning and produced five models (Fig. 1), all of which were implemented through the keras framework of python3.

The VGG16 architecture was selected to transfer learning and its weights were fine-tuned with a low learning rate. VGG16 with 13 convolutional layers and 3 fully connected layers. The first time, after two convolutions of 64 convolution kernels, one pooling is used. The second time, after two 128 convolution kernel convolutions, pooling is used. Repeat the convolution with three 512 convolution kernels twice, and then pooling. Finally, after three full connections. Because it uses a 3×3filter with a stride of 1 to construct the convolutional layer, the padding parameter is the parameter in the same convolution. Then use a 2×2 filter with a stride of 2 to construct the pooling layer, so that the VGG16 network does not have so many hyperparameters. Therefore, one of the advantages of the VGG network is that it does simplify the structure of the neural network, but it is very deep and can learn more deep features.

At the same time VGG16 model had pre-trained on the ImageNet dataset containing 1000 classes which already had learned features that are relevant to our classification issues. The details are as follows (Fig. 1), Model 1: Load the weights of pre-trained VGG16 as a feature extractor and then train full connection layers as their own category classifier; Model 2: loading the weights of pre-trained VGG16 and all layers fine-tuned with a low learning rate; From Model 3 to Model 5: Load the pre-trained VGG16 weights and Model 1 fully-connected layer weights, then freeze different convolutional layers (fixed parameters) and fine-tune the unfrozen layers.

Dense1 layer is a GAP (Global Average Pooling) layer with dropout. GAP was first proposed in and is considered a new technology that can replace the fully connected layer, especially in transfer learning[28, 29]. Assuming that the final output of the convolution layer is a three-dimensional feature map of h × w × d, the specific size is 6 × 6 × 3, after GAP conversion, it becomes an output value of size 1 × 1 × 3, that is, each of the layers h × w will be averaged to a value. There are two advantages, one is that GAP is more simple and natural to convert between the feature map and the final classification. The second is that unlike the FC layer that requires a lot of training and tuning parameters, reducing the spatial parameters will make the model more robust and better resist overfitting. Using our data set, comparing the GAP, GMP and FC layers, we found that the GAP technology was relatively stable in accuracy and loss rate, and the effect of anti-overfitting was obvious. Dropout technology was first proposed in [30]. Later, it was applied to CNN network models for image classification [31]. The effect of anti-overfitting is obvious. The principle is to discard some neurons with probability P during transmission, other neurons are retained with probability q = 1-

p, and the output of the discarded neurons is set to zero [32]. In this article, P was chosen to be 0.2(0.1,0.2,0.5 were tested). Dense2 layer is a full connection layer with 100 neurons and ReLu as activation function.Dense3 layer using Softmax function with two neurons achieving two classification effect.

All images were resized into 224×224 then input into the per-trained model (Model 1 to Model 5). The hyper parameters of the model were selected: learning rate to 0.0001(0.01, 0.001 were tested), batch size to 64, number of epochs to 20, loss was categorical_crossentropy, Adam optimizer (Rmsprop, SGD, Adam were tested).In order to prove the effectiveness of the designed models in this paper, a CNN model with three convolution layers with a ReLU activation and followed by max-pooling layers to train it from scratch as a benchmark for comparison. All images were resized into 224×224 then input into the CNN. The hyper parameters were: learning rate to 0.001, batch size to 32 number of epochs to 100, Adam optimizer, loss was categorical_crossentropy.

D. Model Evaluation

The confusion matrix is an indicator to judge the result of the model, and is generally used to judge the quality of the classifier [33]. The confusion matrix can reflect the effectiveness and accuracy of the model in more detail than the evaluation index. We can clearly see the identification of each type of sample. Here the good category was defined as the label 0 and the bad category as the label 1. For the label 0 : the number of good category photos correctly identified as label 0 by the model is a, the number of label incorrectly identified as label 1 is b; For the label 1: the number of correctly identified as label 1 by the model is c, and the number of incorrectly identified as label 0 by the model is d. Therefore, the form of confusion matrix as follow:

$$\text{Confusion Matrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{4}$$

The following evaluation indicators are all based on the confusion matrix, specific calculation process is as follows:

- In the equation(5) and (6), precision (Accuracy rate) means the proportion of the good(bad) category photos that are actually classified as label 0(1). In general, the higher the accuracy, the better the model.

$$\text{Precision}(0) = \frac{a}{a+c} \tag{5}$$

$$\text{Precision}(1) = \frac{d}{b+d} \tag{6}$$

- In the equation(7) and (8), recall represents the ratio of the number of good(bad) photos that the model identified as label 0(1) to the total number of good(bad) category dataset. In general, the higher the Recall, the more good(bad) photos are predicted by the model.

$$\text{Recall}(0) = \frac{a}{a+b} \tag{7}$$

$$\text{Recall}(1) = \frac{d}{c+d} \tag{8}$$

- In the equation (9) and (10), F1_Score which is defined as the harmonic average of correctness and recall. The value of F1-Score is from 0 to 1, with 1 being the best and 0 being the worst.

$$F1_{Score(0)} = \frac{2Precision(0)*Recall(0)}{Precision\ (0)+Recall(0)} \tag{9}$$

$$F1_{Score(1)} = \frac{2Precision\ (1)*Recall(1)}{Precision(1)+Recall(1)} \tag{10}$$

- The abscissa of the ROC curve is the false positive rate (FPR is equal to recall(1)) and the ordinate is the true positive rate (TPR is equal recall(0)). AUC (Area under Curve) is defined as the area under the ROC curve, which clearly indicates which classifier has the better effect [31]. The classifier with larger AUC is better: 0.5 <AUC <1, which is better than random guessing. If the classifier properly sets the threshold, it can have predictive value; AUC = 0.5, like the follower guess, the model has no predictive value; AUC <0.5, which is less accurate than random guessing [34].

E. Applicability of the Model to Other Data Sets

Forty photos in the paper[17] were analyzed by the method described in [8]. 14 photos grouped as bad category which were “Blackboard”, “Classroom”, “Textbook”, “Dorm room”, “School”, “School bag”, “Diploma”, “Coal”, “Rice balls”, “Dumplings”, “Spring Festival Couplets”, “Zhongshan Mausoleum”, and “Ping pong”.

There were 16 photos grouped as good category, “Popcorn”, “Sausage”, “Ravioli”, “Meatballs with soy sauce”, “Fried rice”, “Hot pot”, “Spring Festival Gala”, “New Year”, “Dragon Boat Race”, “Children's day”, “Black and white TV”, “Tea bottle”, “Vintage bicycle”, “Kerosene lamp”, “Vanishing cream”, and “Quilt”. Fig. 4 shows the distribution of photos. All of them were inputted into the model for prediction to evaluate how well the model for other data.

F. Establishment of the Photo Conversation Support System

Finally a preliminary intergenerational photo conversation support system was established as shown in Fig. 2.

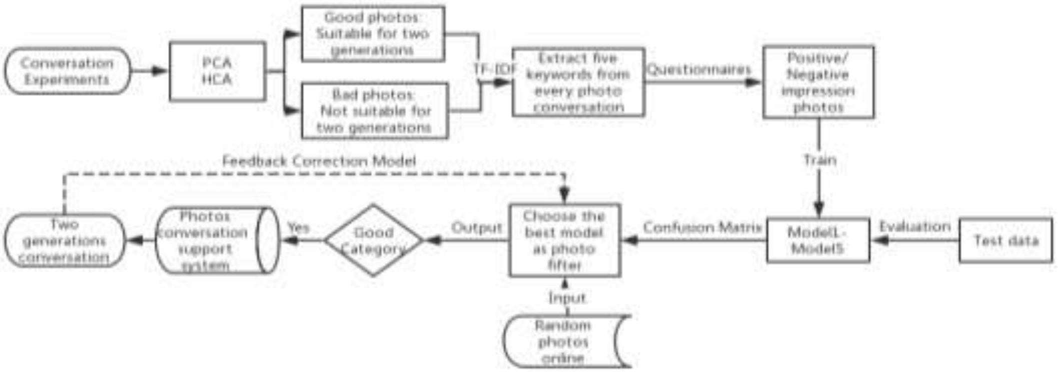
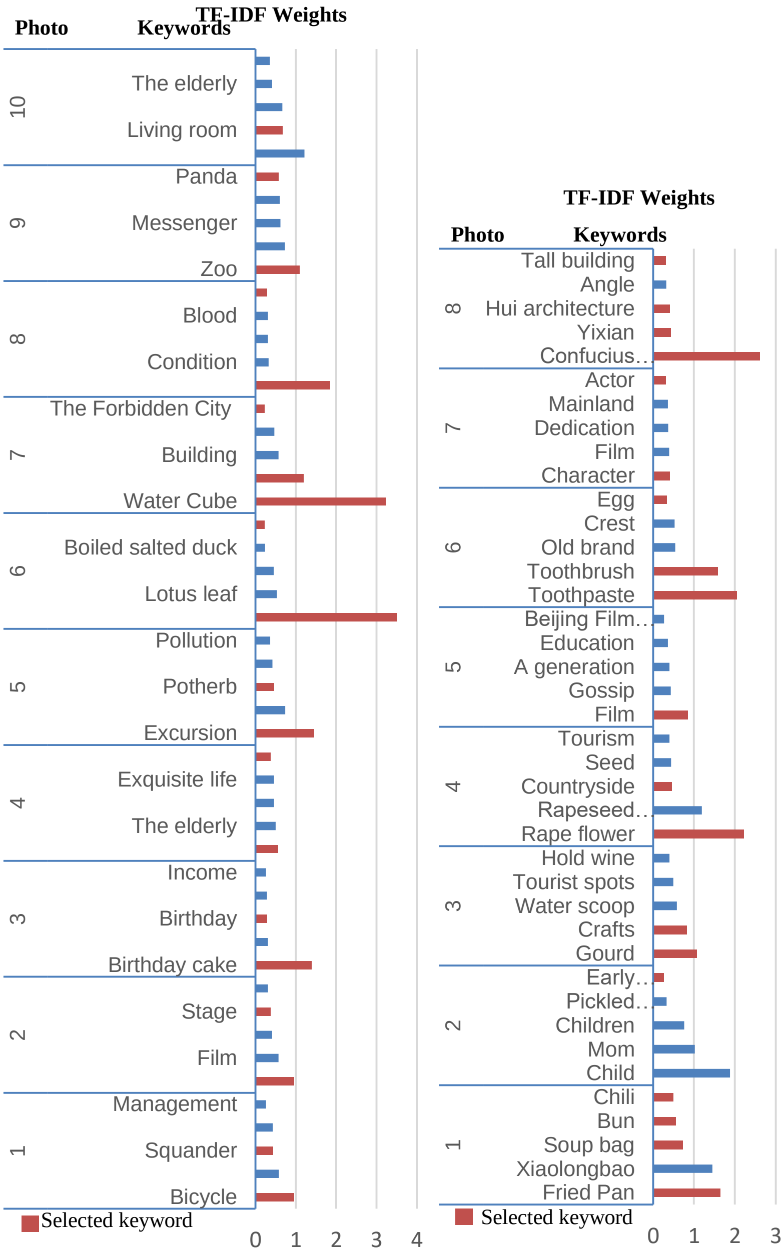

Conversation Experiments
PCA HCA
Good photos: Suitable for two generations
Bad photos: Not suitable for two generations
TF-IDF
Extract five keywords from every photo conversation
Questionnaires
Positive/ Negative impression photos
Train
Feedback Correction Model
Two generations conversation
Photos conversation support system
Yes
Good Category
Output
Choose the best model as photo filter
Confusion Matrix
Model1-Model5
Evaluation
Test data
Input
Random photos online

Fig. 2. Photo Conversation Support System.



(a) Positive Impression Photos

(b) Negative Impression Photos

Fig. 3. The Result of Keyword Extraction and Questionnaires.

IV. Results

A. Dataset Division

Records of conversations for each photo were extracted top five keywords by the TF-IDF technology. Then through questionnaires, each group selected twenty keywords by the elderly and the youngers. As shown in Fig. 3, every selected keyword had seventy-five photos selected by the young and the elderly.

Therefore, we have two photo categories: those with a positive impression (good category) and those with a negative impression (bad category), each with 1500 photos. Next, the photos in each category were randomly divided into a training set and a validation set at a 2: 1 ratio. The training set was 1000 photos and the validation set was 500 photos per category. The original 18 pictures were used as a test set, ten of which were in the good category and eight of which were in the bad category.

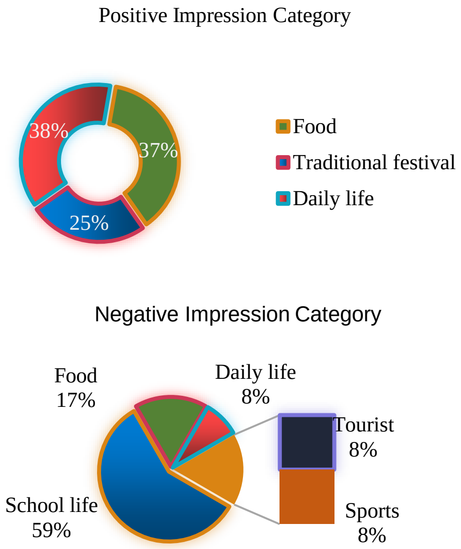


Fig. 4. The Distribution of Photo Categories.

B. Photo Classification Model

Table II shows that transfer learning and fine-tuning the network is more effective than training from scratch (CNN) for small data sets. The accuracy rate increased from 80% to over 94%. Compared to Model 1 using VGG16 as a feature extractor and training a fully connected layer as a classifier, Models 2 to 5 fine-tuned the VGG16 network to obtain better results. The best result was the accuracy rate increasing from 94.8% to 99.8%. And Model 3 and Model 4 have the highest accuracy on the training set and validation set. However, it is not important whether it identifies good category photos as bad category photos, as it will not affect our selection of photos for dialogue. The important thing is that we must not identify bad category photos as good category photos, so a higher rate of bad classification recognition was needed. To achieve this, the confusion matrix of the validation set and the test set used to further evaluate the model (Table III). From the results in Table III, it shows that Model 3 and 4 had the highest recognition rates (the recall value of the bad category) for the test set, but in the Validation set Model 4 was better than Model 3 (the classifier with larger AUC is better). Therefore Model 4 was finally selected.

C. Optimization of Model Parameters

In order to further improve the model, we increased the number of neurons in the Dense2 layer (Table IV). We found that in Model 4 every bad category photo was recognized when we applied a Dense2 layer with 150 neurons.

D. Applicability of the Model to Other Datasets

From Table V, we can see that for the data set in paper [17], the recognition rate of Model 4 with 150 neurons for bad photos is still acceptable as most of them can still be detected. The AUC value of Model 4 is greater than 0.5, so it has predictive value.

TABLE II. The Accuracy and Loss of Training and Validation Data

Model	Accuracy	Loss	Val Accuracy	Val loss
CNN	0.8020	0.4702	0.7160	0.5907
1	0.9480	0.1298	0.9140	0.2868
2	0.9945	0.0264	0.9300	0.1998
3	0.9980	0.0069	0.9560	0.1350
4	0.9980	0.0056	0.9560	0.1298
5	0.9500	0.1374	0.9320	0.2198

TABLE IV. RESULTS OF MODEL EVALUATION

Model	**Data set**	**Label**	**Confusion matrix**	**Precision**	**Recall**	**F1**	**ROC AUC**
1	Validation	Good 0	$\begin{vmatrix}471 & 29\\ 54 & 446\end{vmatrix}$	0.90	0.94	0.92	0.917
		Bad 1		0.94	0.89	0.91	
	Test	Good 0	$\begin{vmatrix}8 & 2\\ 2 & 6\end{vmatrix}$	0.80	0.80	0.80	0.775
		Bad 1		0.75	0.75	0.75	
2	Validation	Good 0	$\begin{vmatrix}471 & 29\\ 29 & 471\end{vmatrix}$	0.94	0.94	0.94	0.942
		Bad 1		0.95	0.94	0.94	
	Test	Good 0	$\begin{vmatrix}10 & 0\\ 1 & 7\end{vmatrix}$	0.91	1.00	0.95	0.938
		Bad 1		1.0	0.88	0.93	
3	Validation	Good 0	$\begin{vmatrix}488 & 12\\ 40 & 460\end{vmatrix}$	0.92	0.98	0.95	0.948
		Bad 1		0.97	0.92	0.95	
	Test	Good 0	$\begin{vmatrix}10 & 0\\ 1 & 7\end{vmatrix}$	0.91	1.0	0.95	0.938
		Bad 1		1.0	0.88	0.93	
4	Validation	Good 0	$\begin{vmatrix}477 & 32\\ 57 & 443\end{vmatrix}$	0.94	0.95	0.95	0.949
		Bad 1		0.95	0.94	0.95	
	Test	Good 0	$\begin{vmatrix}10 & 0\\ 1 & 7\end{vmatrix}$	0.91	1.0	0.95	0.938
		Bad 1		1.0	0.88	0.93	
5	Validation	Good 0	$\begin{vmatrix}475 & 25\\ 57 & 443\end{vmatrix}$	0.89	0.95	0.92	0.919
		Bad 1		0.95	0.89	0.92	
	Test	Good 0	$\begin{vmatrix}7 & 3\\ 2 & 6\end{vmatrix}$	0.78	0.70	0.74	0.725
		Bad 1		0.67	0.75	0.71	

TABLE V. SELECTION OF THE NUMBER OF NEURONS IN THE DENSE2 LAYER

Dense2	**Date sets**	**Label**	**Confusion matrix**	**Precision**	**Recall**	**F1**	**ROC AUC**
150 Model 4	Validation	0	$\begin{vmatrix}485 & 15\\ 27 & 473\end{vmatrix}$	0.95	0.97	0.96	0.958
		1		0.97	0.95	0.96	
	Test	0	$\begin{vmatrix}9 & 1\\ 0 & 8\end{vmatrix}$	1.0	0.9	0.95	0.950
		1		0.89	1.0	0.94	
256 Model 4	Validation	1	$\begin{vmatrix}480 & 20\\ 28 & 472\end{vmatrix}$	0.94	0.96	0.95	0.952
		0		0.96	0.94	0.95	
	Test	1	$\begin{vmatrix}10 & 0\\ 1 & 7\end{vmatrix}$	0.91	1.00	0.95	0.938
		0		1.0	0.88	0.93	

TABLE VI. RESULTS OF NEW TEST DATA BY BEST MODEL

Dense2	**Date set**	**Label**	**Confusion matrix**	**Precision**	**Recall**	**F1**	**ROC AUC**
150 Model 4	Test	1	$\begin{vmatrix}8 & 8\\ 4 & 10\end{vmatrix}$	0.67	0.50	0.57	0.607
		0		0.56	0.71	0.63	

VI. Discussion

1) Although the data sets of this paper were extracted from the conversation content of the other paper, it is clear that the keywords selected in Fig. 3 were similar to Fig. 4 (the distribution of paper [17]). At the same time, the training set and validation set are obtained by extracting keywords through TF-IDF technology, but accuracy of the test set (original photos) is very good. From both aspects proved that the TF-IDF technology is effective for expanding the data set from the content of the photos.

2) The selected positive impression photos were closely related to the daily life of the elderly, for example “famous attractions” (the elderly will travel with tour groups after retirement), “birthdays” (three generations celebrating their birthdays together), “zoos” (for visiting with their grandchildren). The negative impression pictures are all no longer related to their life, such as “toothpaste and toothbrush” (many elderly people wear dentures), “school life” (because many people did not have the opportunity to go to school in their youth), “sports” (they cannot maintain their physical strength) and so on. This phenomenon provides a general direction for us to choose photographs of conversations between two generations in the future.

3) To solve overfitting, the general approach is data augmentation [23]. As indicated on Table III in the CNN model, the accuracy of the validation set is almost 10% worse than the accuracy of the training set. The cause is that the expanded samples are still highly correlated. Therefore, for small data sets, it is not enough to consider only data amplification, and the model must be improved at the same time.

4) The deeper model VGG16 were chosen to transfer learning and fine-tuning , Table IV shown that the accuracy of the verification set and the training set similar and the accuracy of both increased to over 91%. The reason for this result is that models provided in this paper reduced the irrelevant information of photos. From the perspective of models, the main focus of preventing overfitting should be the entropy capacity of the model, or how much information the model can store [25, 27]. There are different ways to modulate entropic capacity, the main one is the choice of the number of parameters in your model, i.e. the number of layers and the size of each layer [35]. Although a model that can store more information may become more accurate by using more convolutions, there is a risk of overfitting to store content that is not related to classification features. Therefore more deeper model chosen with using transfer learning and fine-tuning to reduce the weights is necessary for small data sets.

5) Model 2, which fine-tuned the entire model, was superior to Model 1, which only trained the classifier. However, Model 2 was less accurate than Model 3 and Model 4. This was because Model 2 does not load the already trained classifiers from Model 1 like Model 3 and Model 4, but instead directly initializes a fully connected layer randomly on top of a pre-trained convolutional base layer. It leaded to large gradient updates triggered by the randomly initialized weights which wrecked the learned weights in the convolutional base.

Model 3, Model 4, and Model 5 are all loaded with the classifier weights already trained in Model 1. Among them, the accuracy of Model 3 and Model 4 is the same in the training set and the validation set, but Model 4 is better than Model 3 in the test set. Because Model 3's entire network had a very large entropic capacity and thus a strong tendency to overfitting.

The anti-overfitting effect of Model 5 is better than Model 3 and Model 4. The validation set is only 2% worse than the training set, but it performs the worst on the test set. Although the features learned by low-level convolutional blocks are more general and less abstract than those found higher-up, Model 5 only trains the last block (more specialized features) so is not enough for our data. Model 4 trained the last two blocks so is more effective for our data.

6) Related researches [1][16][17]has not verified their classification results with other data sets. But this paper used other data sets to verify its applicability to others. At the same time, [17] conversation experiment was done in Suzhou, [1] conversation experiment was done in Nanjing. This paper contains a questionnaire done in Shanghai. In the end, it was found to be inaccurate in recognizing photos in the good category, which shows that peoples' favorite things are usually highly personal. The model recognizes the bad category better. The first reason for this is because the photos with negative impressions are often something that has not been used recently by elderly people, such as toothpaste, toothbrushes, etc. Second, because the three locations tested belong to different cities in the south, it was easy to identify the differences in eating and cultural habits, such as not liking dumplings that are preferred by northerners. Therefore, in the future, we should avoid choosing photos that are not suitable for the conversation between the two generations from these two aspects.

VII. Conclusion and Perspectives

1) Using TF-IDF technology to extract keywords and then using a questionnaire to select photos provides is effective for expanding the data set based on the content of the photos.

2) Although Model 4 has a low recognition rate for selecting photos that are suitable for intergenerational conversation , it can effectively filter out photos that are not suitable for intergenerational conversation. Finally our goal is finished that this system filtered out photos that are unsuitable for conversation and kept the suitable photos.

3) This system was effective in the three cities of Suzhou, Shanghai and Nanjing, indicating that the system is likely to be applicable to more regions. Therefore, the next step is to select photos via the initially established photo dialogue support system, go to other cities for intergenerational dialogue, and then modify Model 4. So that this system can be applicable to most regions.

4) This system has some flaws, while select suitable photos for Intergenerational Conversation more factors should be considered, such as different age groups, different educational backgrounds, different hobbies, different personalities, etc. Later more models need be built to learn these characteristics.

ACKNOWLEDGMENT

This research was supported by JSPS KAKENHI grant number 15H01698. We would also like to express our gratitude to the China Scholarship Committee (CSC) and to all the young and the elderly who assisted in completing the experiment.

REFERENCES

[1] X. Dong, “Advances in the study of elderly care models in China,” Chinese Journal of Gerontology, vol. 039, no. 004, pp. 996-999, 2019.

[2] Y. D. Ya Zhu, Changqing Wang, “Research status and innovative thinking of healthy old-age care model,” Journal of Nanjing Medical University (Social Science Edition), vol. 018(002), pp. 103-106, 2018.

[3] S. He, “Research on the Mutual Assistance and Cooperative Social Pension Model from the Perspective of Developmental Welfare,” Rural Economy, no. 01, pp. 73-76, 2014.

[4] B. Li, and C. Jiang, “The Experience and Enlightenment of Community Pension Model in Britain and Japan,” Foreign trade, no. May 2015, pp. 58-59, 2015.

[5] Q. Wang, “Nursing service system and countermeasures for the elderly living alone and empty-nesters,” China Economic Times, vol. 12851, April 2019

[6] J. Liu, “A renew of research on the mental health of urban empty nest elderly in China,” Chinese Journal of Nursing, vol. 043, no. 5, pp. 457-459, 2008.

[7] Y. Xue, C. Zhang, H. Zhao, X. Zheng, and Y. Cai, “Depression status and influencing factors of empty nest elderly based on structural equation model,” Chinese Journal of Disease Control, vol. 23, no. 10, pp. 1181-1185, 2019.

[8] Z. Xiaochun, C. Dong-Eun, P. Siriaraya, and N. Kuwahara, “Sentiment Analysis and Classification of Photos for 2-Generation Conversation in China,” International Journal of Advanced Computer Science and Applications, vol. 10, no. 10, 2019.

[9] A. Astell, N. Alm, G. Gowans, M. Ellis, R. Dye, J. Campbell, and P. Vaughan, “Working with people with dementia to develop technology: The CIRCA and Living in the Moment projects,” PSIGE newsletter, vol. 64, 2009.

[10] A. Astell, M. Ellis, N. Alm, R. Dye, J. Campbell, and G. Gowans, “Facilitating communication in dementia with multimedia technology,” Brain and Language, vol. 91, no. 1, pp. 80-81, 2004.

[11] A. J. Astell, N. Alm, G. Gowans, M. P. Ellis, R. Dye, and J. Campbell, “CIRCA: a communication prosthesis for dementia,” Technology and aging, pp. 67-76, 2008.

[12] A. J. Astell, M. P. Ellis, L. Bernardi, N. Alm, R. Dye, G. Gowans, and J. Campbell, “Using a touch screen computer to support relationships between people with dementia and caregivers,” Interacting with Computers, vol. 22, no. 4, pp. 267-275, 2010.

[13] N. Alm, R. Dye, G. Gowans, J. Campbell, A. Astell, and M. Ellis, “A communication support system for older people with dementia,” Computer, vol. 40, no. 5, pp. 35-41, 2007.

[14] B. A. Purves, A. Phinney, W. Hulko, G. Puurveen, and A. J. Astell, “Developing CIRCA-BC and exploring the role of the computer as a third participant in conversation,” American Journal of Alzheimer's Disease & Other Dementias®, vol. 30, no. 1, pp. 101-107, 2015.

[15] D. I. Fels, and A. J. Astell, “Storytelling as a model of conversation for people with dementia and caregivers,” American Journal of Alzheimer's Disease & Other Dementias®, vol. 26, no. 7, pp. 535-541, 2011.

[16] M. Iwamoto, N. Kuwahara, and K. Morimoto, “Comparison of Burden on Youth in Communicating with Elderly using Images Versus Photographs,” International Journal of Advanced Computer Science and Applications, vol. 6, no. 10, 2015.

[17] Z. Xiaochun, M. Iwamoto, and N. Kuwahara, “Evaluation of Photo Contents of Conversation Support System with Protocol Analysis Method,” International Journal of Advanced Computer Science and Applications, vol. 9, no. 4, 2018.

[18] P. B. D. P. A. Vaidya, “Document clustering: TF-IDF approach,” 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, pp. 61-66, 2016.

[19] A. A. H. A. E. K. I. E. M. Galinium, “Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach,” 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, pp. 1-4, 2014.

[20] A. Gulli, and S. Pal, Deep learning with Keras: Packt Publishing Ltd, 2017.

[21] K. B. Ahmed, Jelodar, Ahmad Babaeian, “Fine-Tuning VGG Neural Network For Fine-grained State Recognition of Food Images,” arXiv pre-print server, vol. abs/1809.09529, 2018.

[22] A. M. Dawud, K. Yurtkan, and H. Oztoprak, “Application of Deep Learning in Neuroradiology: Brain Haemorrhage Classification Using Transfer Learning,” Computational Intelligence and Neuroscience, vol. 2019, pp. 1-12, 2019.

[23] Y. Shima, “Image Augmentation for Object Image Classification Based On Combination of Pre-Trained CNN and SVM,” Journal of Physics: Conference Series, vol. 1004, pp. 012001, 2018.

[24] D.-X. Xue, R. Zhang, H. Feng, and Y.-L. Wang, “CNN-SVM for Microvascular Morphological Type Recognition with Data Augmentation,” Journal of Medical and Biological Engineering, vol. 36, no. 6, pp. 755-764, 2016.

[25] O. N. Belaid, and M. Loudini, “Classification of Brain Tumor by Combination of Pre-Trained VGG16 CNN,” Journal of Information Technology Management, vol. 12, no. 2, pp. 13-25, 2020.

[26] K. Rangasamy, M. A. As’ari, N. A. Rahmad, and N. F. Ghazali, “Hockey activity recognition using pre-trained deep learning model,” ICT Express, 2020.

[27] W. Setiawan, M. I. Utoyo, and R. Rulaningtyas, “Transfer learning with multiple pre-trained network for fundus classification,” Telkomnika, vol. 18, no. 3, 2020.

[28] S. H. Kassani, P. H. Kassani, M. J. Wesolowski, K. A. Schneider, and R. Deters, “Breast cancer diagnosis with transfer learning and global pooling,” arXiv preprint arXiv:1909.11839, 2019.

[29] M. Lin, Q. Chen, and S. Yan, “Network in network,” arXiv preprint arXiv:1312.4400, 2013.

[30] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” arXiv preprint arXiv:1207.0580, 2012.

[31] A. Krizhevsky, “ImageNet classification with deep convolutional neural networks,” Communications of the ACM, vol. 60(6), pp. 84-90, 2017.

[32] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” Journal of Machine Learning Research, vol. 15, pp. 1929-1958, 2014.

[33] S. Visa, B. Ramsay, A. L. Ralescu, and E. Van Der Knaap, “Confusion Matrix-based Feature Selection,” MAICS, vol. 710, pp. 120-127, 2011.

[34] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation." pp. 1015-1021.

[35] F. Chollet, “Building powerful image classification models using very little data,” Keras Blog, 2016.

Performance Analysis of Advanced IoT Encryption on Serialization Concept: Application Optimization Approach

Johan Setiawan[1], Muhammad Taufiq Nuruzzaman[2]*
Department of Informatics
Universitas Islam Negeri Sunan Kalijaga Yogyakarta
Yogyakarta, Indonesia

***Abstract*—This study investigates the effect of serialization concepts with cipher algorithms and block mode on structured data on execution time in low-level computing IoT devices. The research was conducted based on IoT devices, which are currently widely used in online transactions. The result of overheating on the CPU is fatal if the encryption load is not reduced. One of the consequences is an increase in the maintenance obligations. So that from this influence, the user experience level will have bad influence. This study uses experimental methods by exploring serialization, ciphers, and block mode using benchmarks to get better data combination algorithms. The four test data groups used in benchmarking will produce an experimental benchmark dataset on the selected AES, Serpent, Rijndael, BlowFish, and block mode ciphers. This study indicates that YAML minify provides an optimal encryption time of 21% and decryption of 27% than JSON Pretty if an average of the whole test is taken. On the other hand, the AES cipher has a significant effect on the encryption and decryption process, which is 51% more optimal for the YAML minify serialization**

***Keywords*—Internet of Things; benchmark; cipher; block mode; serialization**

I. Introduction

Computers and IoT are useful in assisting the activities of certain individuals or business groups. These business domains include Trade, Transportation, Health [1]–[4], and other specific matters discussed in research [5]–[7]. With the expansion of computers, all devices equipped with a microprocessor have been embedded to sustain mobility and the device's toughness. Devices controlled automatically or remotely are a family of IoT (Internet of Things) supporting devices. IoT uses the M2M (machine-to-machine) concept communication[8] without any human relationship [2][9].

Communication between IoT devices uses information data and instructions that have been designed or regulated by the manufacturer. The information sent and received by devices usually does not want to be known or understood by parties or devices [10][4][11][12] with no interest in destroying or converting the information. Therefore, manufacturers should consider durability and safety at a low cost [13]. The information security risks can be in the form of modifications or interruptions, and these risks can affect the continuity of the process or business flow that is running[8][12][14]. In tackling these threats, data encryption is required [15]. Encryption is a method used to convert original data into artificial data to become rugged and not accessible for humans to read. The encryption process's drawback tends to impose more processing on the microprocessor embedded in an IoT device. It can result from small and limited microprocessor capabilities [16][17] and large amounts of data in the encryption process [18]–[20]. As a result of an encryption algorithm's complexity, the microprocessor on the IoT device is more burdened.

The direct effect of microprocessor devices that get high loads or pressures to overheat is the length of the computation process of a device so that it affects UX (User Experience) because it can reduce the level of efficiency [21][22] Users will feel bored in waiting for computation so that it has an impact on ongoing business processes [21][22][23][24]. On the other hand, the impact of overheating microprocessor pressure is that the device is not durable. It harms device providers that have to carry out more routine maintenance. In [3] research discussed one method of encrypting data with a Catalan object base and two structural combinations on IoT devices; however, that study did not discuss the concept of serialization in structured data to the encryption process. Therefore, The research related to the analysis and evaluation of several algorithms that are often used in data encryption which includes; AES, Rijndael, Serpent and Blowfish. The encryption process uses several different data serialization concepts to improve application performance on IoT systems thus that it can provide less computation, time and memory and provide a better impact on user UX (User Experience) while sustaining a level of security information. At the same time, the advantages for companies that will be gained from this article are used as an option for IoT device providers in dealing with the problem of overheating on the low-level computational microprocessor utilized.

The rest of this article is organized as follows. Section II discusses previous research that has been carried out as a literature study for the author. Section III discusses the research methods used. The experiment is a method used by the author to obtain benchmark data on a combination of serialization, cipher and block mode. Section IV provides data design, benchmark flow design, data collection process and analysis process from experiments. Section V concludes this article.

*Corresponding Author

II. Related Works

Data is an essential thing in business that must be secured when transmitted over public networks. It relates to attacks that threaten data modification and interruption. Encryption and data authentication schemes that are implemented can shield data from these attacks to not be read by unauthorized people [4][10]. Nevertheless, on the other hand, encryption can burden the microprocessor [3], resulting in overheating. Sudip Maitra et al. Have published research related to evaluating the performance of the encryption algorithm on IoT devices with the XTEA and AES algorithms to obtain an algorithm with more optimal memory, time, and energy. The research has been conducted using an experimental method utilizing an oscilloscope device to help identify the energy consumed in the encryption process. From these results, it was found that the XTEA algorithm is better in terms of efficiency if the IoT device does not use the AES accelerator [18]. On the other hand, Geovandro C. et al. Has researched the evaluation of cryptographic algorithms' performance on IoT and operating systems [17].

Nur Rachmat et al. has implemented the analysis of the performance of Rijndael, Serpent, Twofish on an android smartphone. The conclusion of the research that Serpent has good performance at execution time [25]. Furthermore, Muzafer H. Saračević et al. have been carried out providing a proposal to use encryption on Catalan object-based IoT and two structural combinations. In this research, the whole procedure is based on the Catalan number's characteristics and the representation numbers and combinatorial problems. Apart from improving the quality of encryption, that study recommended lightweight computing like e-health and smart cities [3]. Moreover, several studies are almost similar in [26]–[28].

III. Research Method

Researchers use the research method to collect data that followed up as material for investigation and analysis. The research method provides information from the research design to be composed of time, place, and data source. At this condition, the researcher applies experimental research methods. It will provide an overview of the effect of data serialization of pretty JSON, JSON minify, YAML, and YAML minify as machine-to-machine communication. It is against several encryption algorithms/ciphers at AES, Blowfish, Rijndael, and Serpent on IoT devices to get better device performance. The steps taken in this study were coding, data collection, data grouping, benchmarking, and the analysis process. The procedure of the research method used in this study can be seen in Fig. 1.

A. Research Tools

Tools are vastly crucial in research since it can affect the results of the analysis to be performed. This study uses an IoT device in the form of Orange Pi Zero. It is a tiny computing device with a more complex system to assist the data collection process with specifications shown in Table I.

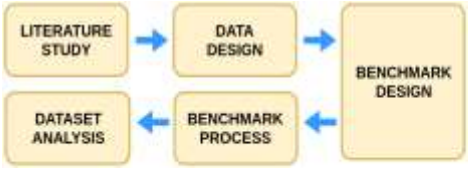


Fig. 1. Research Method.

TABLE I. IoT Hardware Specification

Type	Specifications
CPU Manufacture	AllWinner / ARMv7
CPU Core	4 Cores
CPU Speed	1.5 GHz
RAM	512 MB
Disk	32 GB
OS	Linux Ubuntu 18.04 LTS

B. Data Collection

Data will operate as serialized data in JSON and YAML with the pretty and minify schemes in this process. From some of these data serialization concepts, the next step is to benchmark the data encryption process to adjust the system or algorithm to get better processing on specific platforms [29]. The benchmarking process in this study uses the Golang programming language. The author uses that in benchmarking to get data benchmarks with low-level programming languages [30]. Thus the golang application will be compiled into binary so that it becomes faster [31].

C. Data Grouping

The grouping process will determine the percentage level of time efficiency or processing carried out in the encryption and decryption process. The process will group on the type of encryption or decryption used based on data serialization. From the benchmark results obtained, researchers can compare these results. A more efficient variety of data serialization, cipher, and block mode combination will be discovered for the IoT application encryption and decryption process.

IV. Result and Discussions

A. Benchmark Design

Several schemes of the data structure to be used in the benchmark have different characteristics so that the data used will have various levels of influence. The data sample is traditional data commonly used in communication between devices. The traditional JSON data structure is used as a variety of file sizes, including 1.5 KB, 2.1 KB, 3.0 KB and 3.5 KB. The data will be converted with serialization concept.

At this step, some of the cipher and block mode algorithms listed in Table II will be designed as a benchmark process stage executed after the data serialization process. In this case, it is assumed that the data have been through the serialization process such as JSON and YAML with the concept of minifying and pretty. The process diagram can be seen in Fig. 2.

TABLE II. CIPHER AND BLOCK MODE

Cipher	Block Modes
AES-256	CBC, CTR, ECB
BlowFish	CBC,EBC,OFB
Serpent	CBC,ECB,OFB,NOFB,CFB,NCFB,CTR
Rijndael-256	CBC,ECB,OFB,NOFB,CFB,CTR,NCFB

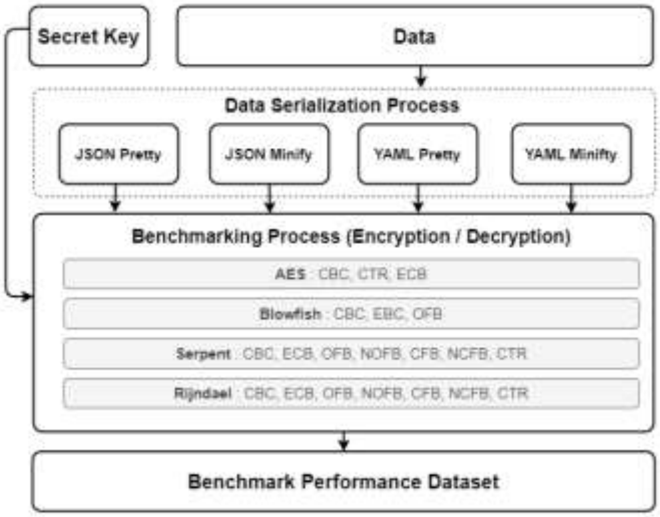

Secret Key
Data
Data Serialization Process
JSON Pretty
JSON Minify
YAML Pretty
YAML Minifty
Benchmarking Process (Encryption / Decryption)
AES : CBC, CTR, ECB
Blowfish : CBC, EBC, OFB
Serpent : CBC, ECB, OFB, NOFB, CFB, NCFB, CTR
Rijndael : CBC, ECB, OFB, NOFB, CFB, NCFB, CTR
Benchmark Performance Dataset

Fig. 2. Benchmarking Flow.

In Fig. 2, there are data arrays and secret keys that have been provided. The data is used in the serialization process and converted into an appropriate data structure. Furthermore, in the benchmarking process section, the data used is data that has gone through serialization. In conducting benchmarks, serialized data will be looped with the number of *n* in the benchmark function algorithm and block mode using the provided secret key. After the benchmarking is complete, the performance dataset will be used to conduct research using comparative analysis.

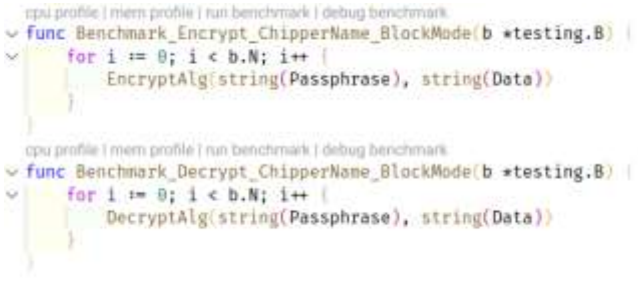
In Table III, Schemes 1 and 2 are schemes that are often operated in serialized data encryption and usually used in applications. Meanwhile, Schemes 3 and 4 are comparison schemes using different data serialization concepts and rarely used from traditional types. With these functions' combination, the benchmark process will be carried out to gain speed/application optimization on the IoT encryption process.

B. Experiment Flow Design

In directing experiments, the Package Benchmark used by the researcher used a package that was already installed in the Golang programming language. In contrast, the operating system's encryption package is libmcrypt-dev, which is also embedded in the Operating System used to have a better effect on library performance. Furthermore, the encrypt library used in the project is able to be seen in the repository (*https://github.com/mfpierre/go-mcrypt*). In preparing the benchmark function, the researcher provides a specimen that can be seen in Fig. 3.

TABLE III. COMBINATION OF ENCRYPT AND SERIALIZATION

Schemes	Functions
Scheme 1	EncryptAlg(JSON(Data))
Scheme 2	EncryptAlg(Minify(JSON(Data)))
Scheme 3	EncryptAlg(YAML(Data))
Scheme 4	EncryptAlg(Minify(YAML(Data)))


cpu profile | mem profile | run benchmark | debug benchmark

```
func Benchmark_Encrypt_ChipperName_BlockMode(b *testing.B) {
    for i := 0; i < b.N; i++ {
        EncryptAlg(string(Passphrase), string(Data))
    }
}
```

cpu profile | mem profile | run benchmark | debug benchmark

```
func Benchmark_Decrypt_ChipperName_BlockMode(b *testing.B) {
    for i := 0; i < b.N; i++ {
        DecryptAlg(string(Passphrase), string(Data))
    }
}
```

Fig. 3. Benchmark Function Scheme.

In Fig. 3, there is a function with the prefix 'Benchmark' which serves as a marker that the function is applied to perform benchmarking tests against "ChipperName" and "BlockMode." The 'testing' package determines the number of iterations performed in the function. At the same time, 'EncryptAlg' is a function been arranged as an application helper according to the list in Table II and has been adjusted in Table III assuming the 'Data' variable has been serialized first or in 'DecryptAlg' variable 'Data' is data that have been encrypted.

C. Benchmark Result

The purpose of the benchmark process is to find the average execution time of each algorithm on the cipher and block mode used in encryption. The results of this benchmark will then be analyzed in the data analysis step. The formula for the benchmark calculation of all sample data is as follows.

$$f(x, y) = \sum_{k=1}^{n} x\big(y(Sk)\big) \tag{1}$$

Where in that function, *x* is the benchmark function used in Fig. 3. *y* is the data serialization function. *n* is total sample data, in this case, the researcher uses four data samples that have been described, and *S* is the list of sample data used and has been used to serialize. From the results of calculations using the formula 1, the data to be obtained will be listed as in Tables IV and V on each scheme, cipher and block mode used.

In Table IV and V to measure the percentage of time reduction from scheme 1 with scheme 4 uses the formula which can be seen in formula 2 - 4.

$$j(x) = \sum_{k=1}^{n} x\big(b(Sk)\big) \tag{2}$$

$$y(x) = \sum_{k=1}^{n} x\big(m(Sk)\big) \tag{3}$$

$$f(x) = \frac{j(x) - y(x)}{j(x)}\ x\ 100 \tag{4}$$

TABLE IV. BENCHMARK RESULT OF THE ENCRYPT COMBINATION WITH SERIALIZATION

No.	**Cipher**	**Block Mode**	**Total Time (*ns*)**				**Reduction Time from Scheme 1 to 4 (%)**
			Scheme 1	***Scheme 2***	***Scheme 3***	***Scheme 4***	
1	AES-256	CBC	1253844	667272	680424	628043	49,91
2	AES-256	CTR	1805709	1005851	1008796	979993	45,73
3	AES-256	ECB	768435	388808	394922	358955	53.29
4	Blowfish	CBC	4034594	3782520	3778960	3752085	7.00
5	Blowfish	EBC	4130058	3887133	3907471	3867405	6.36
6	Blowfish	OFB	6331005	5072219	5072810	4970101	21.50
7	Serpent	CBC	3108647	2831654	2806282	2773793	10.77
8	Serpent	ECB	3251809	3589981	2934591	2886946	11.22
9	Serpent	OFB	10200276	5750340	6385557	6067675	40.51
10	Serpent	NOFB	3501421	3661894	3163832	3116317	11.00
11	Serpent	CFB	9883248	5502608	6062519	5742311	41.90
12	Serpent	NCFB	3439166	3057695	3118968	3055999	11.14
13	Serpent	CTR	3293048	2998750	2973943	2931130	10.99
14	Rijndael-256	CBC	3526985	3146207	3130867	3071622	12.91
15	Rijndael-256	ECB	3643133	5075309	3245767	3207467	11.96
16	Rijndael-256	OFB	23662065	11162439	13070587	12081937	48.94
17	Rijndael-256	NOFB	3903236	5163250	3452525	3411664	12.59
18	Rijndael-256	CFB	23303105	10860264	12758767	11777678	49.46
19	Rijndael-256	CTR	3702905	3304169	3273811	3232372	12.71
20	Rijndael-256	NCFB	3831561	3252175	3388191	3347123	12.64

TABLE V. BENCHMARK RESULT OF THE DECRYPT COMBINATION WITH SERIALIZATION

No.	**Cipher**	**Block Mode**	**Total Time (*ns*)**				**Reduction Time from Scheme 1 to 4 (%)**
			Scheme 1	***Scheme 2***	***Scheme 3***	***Scheme 4***	
1	AES-256	CBC	1258442	656637	656176	607571	51.72
2	AES-256	CTR	1829366	962004	969786	926403	49.36
3	AES-256	ECB	915100	449850	458863	420494	54.05
4	Blowfish	CBC	4301414	3899693	3900940	3875388	9.90
5	Blowfish	EBC	4363084	3986372	3997765	3956421	9.32
6	Blowfish	OFB	6732089	5171293	5194958	5047958	25.02
7	Serpent	CBC	3418887	2886116	2924647	2867847	16.12
8	Serpent	ECB	3496962	3012961	3029684	2971746	15.02
9	Serpent	OFB	10555775	6454063	6522322	6148184	41.76
10	Serpent	NOFB	3731382	3252595	3264580	3220112	13.70
11	Serpent	CFB	10339117	6133240	6175476	5809008	43.82
12	Serpent	NCFB	3769358	3171058	3176920	3136047	16.80
13	Serpent	CTR	3578474	3054691	3066020	3028263	15.38
14	Rijndael-256	CBC	3850404	3216844	3222877	3172646	17.60
15	Rijndael-256	ECB	3917032	3340315	3345704	3305081	15.62
16	Rijndael-256	OFB	24907847	13150971	13183803	12191129	51.06
17	Rijndael-256	NOFB	4226241	3561972	3577534	3524398	16.61
18	Rijndael-256	CFB	24651813	12773199	12888158	11890058	51.77
19	Rijndael-256	CTR	4037440	3365763	3374559	3335614	17.38
20	Rijndael-256	NCFB	4075122	3500502	3500835	3455867	15.20

In this formula, *j(x)* is a function to get the total benchmark value from the JSON serialization data because *b* is a function used for serializing JSON data. *y(x)* is a function to get the total benchmark value from the YAML serialization since *m* was the YAML minify function. In that formula, the value of *n* is given 4 because this value is the total of the data sample. *S* is a list of sample data used. Thus that in calculating the percentage obtained from the value *f(x)*.

D. Analysis

All benchmark data will be grouped and analyzed more deeply; thus, it becomes more informative data in graphical form. In this study, four data samples have been converted into four types of data structures tested on 20 combinations of cipher and block. In this case, the comparison graph is assumed to be derived from the execution time's total evaluation result. A comparison of the traditional scheme 1 against scheme 4 on the cipher used can be seen in Fig. 4 to 11.

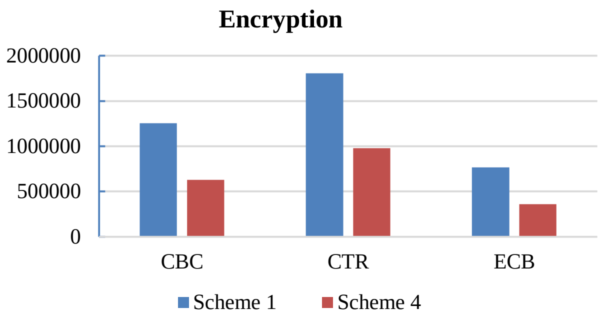


Fig. 4. AES-256 Encryption Time Comparison (*ns*).

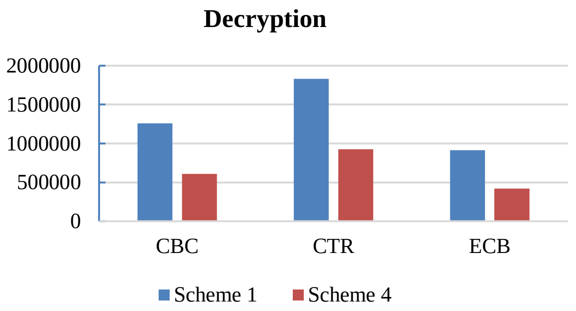


Fig. 5. AES-256 Decryption Time Comparison (*ns*).

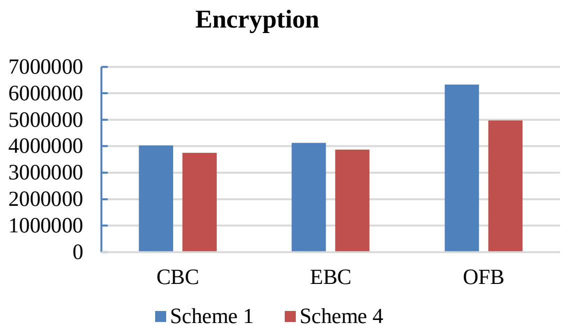


Fig. 6. BlowFish Encryption Time Comparison (*ns*).

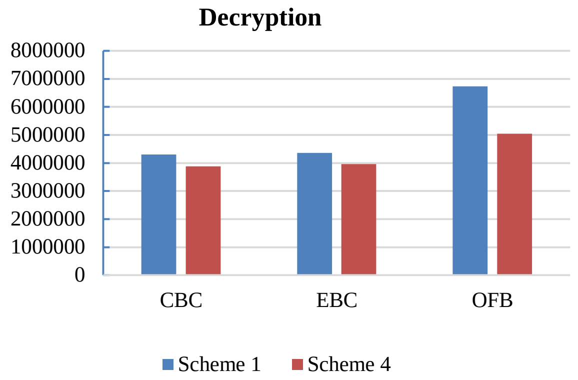


Fig. 7. BlowFish Decryption Time Comparison (*ns*).

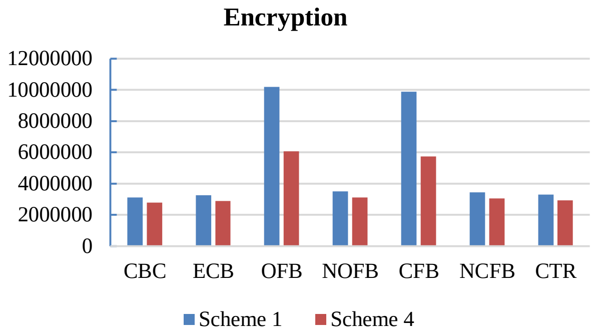


Fig. 8. Serpent Encryption Time Comparison (*ns*).

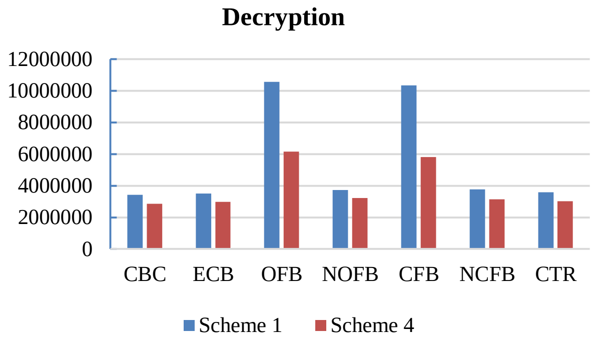


Fig. 9. Serpent Decryption Time Comparison (*ns*).

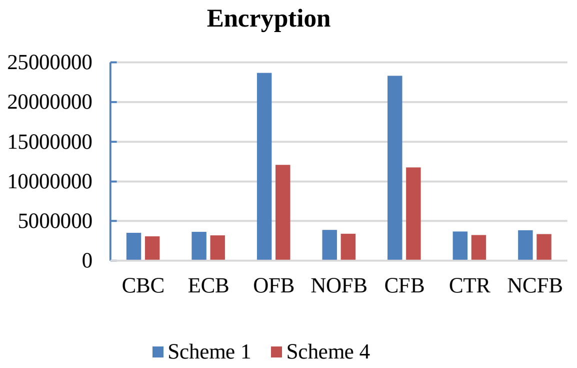


Fig. 10. Rijndael Encryption Time Comparison (*ns*).

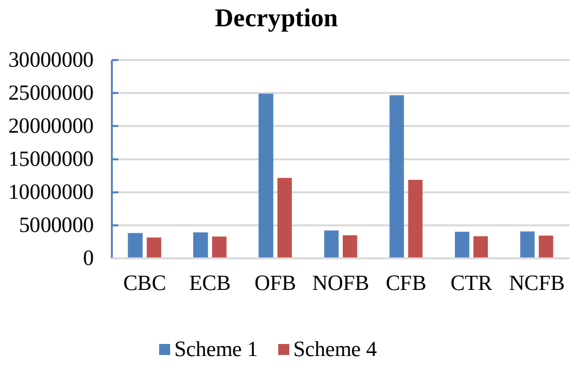


Fig. 11. Rijndael Decryption Time Comparison (*ns*).

Based on Fig. 4 to 11 in the comparison results of schemes 1 and 4 in Table III, the YAML minify serializations method gets an upbeat assessment of the encryption and decryption side. As for the algorithm and block mode used, AES ECB gets better speed than BlowFish, Serpent, and Rijndael. Block modes that provide better performance are ECB, CFB, and OFB. The proposed encryption flow on the low-level IoT platform uses YAML data serialization with the minify scheme. It uses the AES ECB algorithm based on the graphic data described. However, this is very relative to the device used.

V. Conclusion

This research evaluated the cipher and block mode's performance based on several data serialization schemes on low computing devices. The trials' convener carried out using several data serialization. The results were not too significant between scheme one and scheme four on a particular cipher. However, this experiment could have a very significant time-cutting effect on the AES cipher trial with an average of 51% pruning. However, the overall average for encryption will be obtained at 21.85% and 27.36% in decryption. With this research. The hope that it will allow developers to select cipher, block mode, and data serialization to reduce the execution time in the encryption or decryption process.

In the benchmarking process, the author only uses one IoT device. In this case, the author cannot give a definite measure of the figures presented. The author directed a still new system benchmark, and there are no applications that burden the microprocessor. However, it can explain how the serialization, cipher, and block mode combination affects these devices' performance.

The hope for the future, there is further research on this field. For example, by changing data type, data length, protocol, a programming language used or adding the other IoT platform with processor architecture changed.

References

[1] D. Richards, A. Abdelgawad, and K. Yelamarthi, "How Does Encryption Influence Timing in IoT?," 2018 IEEE Glob. Conf. Internet Things, GCIoT 2018, pp. 1–5, 2019.

[2] Y. Hanada, L. Hsiao, and P. Levis, "Smart contracts for machine-to-machine communication: Possibilities and limitations," Proc. - 2018 IEEE Int. Conf. Internet Things Intell. Syst. IOTAIS 2018, pp. 130–136, 2019.

[3] M. H. Saracevic et al., "Data Encryption for Internet of Things Applications Based on Catalan Objects and Two Combinatorial Structures," IEEE Trans. Reliab., 2020.

[4] K. Yelamarthi, M. S. Aman, and A. Abdelgawad, "An application-driven modular IoT architecture," Wirel. Commun. Mob. Comput., vol. 2017, 2017.

[5] H. M. Al-Kadhim and H. S. Al-Raweshidy, "Energy efficient and reliable transport of data in cloud-based IoT," IEEE Access, vol. 7, pp. 64641–64650, 2019.

[6] D. Sharma and D. Jinwala, "Functional encryption in IoT E-Health care system," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9478, pp. 345–363, 2015.

[7] A. D. Dwivedi, L. Malina, P. Dzurenda, and G. Srivastava, "Optimized blockchain model for internet of things based healthcare applications," 2019 42nd Int. Conf. Telecommun. Signal Process. TSP 2019, pp. 135–139, 2019.

[8] P. Radanliev, D. De Roure, C. Maple, J. R. . Nurse, R. Nicolescu, and U. Ani, "Cyber Risk in IoT Systems," Univ. Oxford Comb. Work. Pap. Proj. reports Prep. PETRAS Natl. Cent. Excell. Cisco Res. Cent., vol. 169701, no. 2017, pp. 1–27, 2019.

[9] M. Chen, J. Wan, and F. Li, "Machine-to-machine communications: Architectures, standards and applications," KSII Trans. Internet Inf. Syst., vol. 6, no. 2, pp. 480–497, 2012.

[10] G. A. Utomo, "Ethical hacking," CyberSecurity dan Forensik Digit., vol. 2, no. 1, pp. 8–15, 2019.

[11] M. Stute et al., "A billion open interfaces for Eve and Mallory: MITM, DOS, and tracking attacks on iOS and MACOS through apple wireless direct link," Proc. 28th USENIX Secur. Symp., pp. 37–54, 2019.

[12] I. Fitriani and A. B. Utomo, "Implementasi Algoritma Advanced Encryption Standard (AES) pada Layanan SMS Desa," JISKA (Jurnal Inform. Sunan Kalijaga), vol. 5, no. 3, p. 153, 2020.

[13] M. Weyrich, J. Schmidt, and C. Ebert, "Machine-to-Machine Communication," IEEE, vol. 31, no. 4, pp. 19–23, 2014.

[14] P. Radanliev et al., "Definition of Internet of Things (IoT) Cyber Risk Discussion on a Transformation Roadmap for Standardisation of Regulations Risk Maturity Strategy Design and Impact Assessment," Sensors, no. March, pp. 1–9, 2019.

[15] M. M. Yahaya and A. Ajibola, "Cryptosystem for Secure Data Transmission using Advance Encryption Standard (AES) and Steganography," Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., vol. 5, no. 6, pp. 317–322, 2019.

[16] M. Khari, A. K. Garg, A. H. Gandomi, R. Gupta, R. Patan, and B. Balusamy, "Securing Data in Internet of Things (IoT) Using Cryptography and Steganography Techniques," IEEE Trans. Syst. Man, Cybern. Syst., vol. 50, no. 1, pp. 73–80, 2020.

[17] G. C. C. F. Pereira, R. C. A. Alves, F. L. da Silva, R. M. Azevedo, B. C. Albertini, and C. B. Margi, "Performance evaluation of cryptographic algorithms over IoT platforms and operating systems," Secur. Commun. Networks, vol. 2017, 2017.

[18] S. Maitra, D. Richards, A. Abdelgawad, and K. Yelamarthi, "Performance Evaluation of IoT Encryption Algorithms: Memory, Timing, and Energy," SAS 2019 - 2019 IEEE Sensors Appl. Symp. Conf. Proc., pp. 6–11, 2019.

[19] M. Frustaci, P. Pace, G. Aloi, and G. Fortino, "Evaluating critical security issues of the IoT world: Present and future challenges," IEEE Internet Things J., vol. 5, no. 4, pp. 2483–2495, 2018.

[20] M. Botta, M. Simek, and N. Mitton, "Comparison of hardware and software based encryption for secure communication in wireless sensor networks," 2013 36th Int. Conf. Telecommun. Signal Process. TSP 2013, pp. 6–10, 2013.

[21] A. Crescenzi, D. Kelly, and L. Azzopardi, "Impacts of time constraints and system delays on user experience," CHIIR 2016 - Proc. 2016 ACM Conf. Hum. Inf. Interact. Retr., pp. 141–150, 2016.

[22] D. Biduski, E. A. Bellei, J. P. M. Rodriguez, L. A. M. Zaina, and A. C. B. De Marchi, "Assessing long-term user experience on a mobile health application through an in-app embedded conversation-based questionnaire," Comput. Human Behav., vol. 104, 2020.

[23] O. Noguchi, M. Munechika, and C. Kajihara, "A Study on User Satisfaction with an Entire Operation Including Indefinite-Length Response Time ," Total Qual. Sci., vol. 2, no. 2, pp. 70–79, 2016.

[24] L. T. Yong, "User experience evaluation methods for mobile devices," 2013 3rd Int. Conf. Innov. Comput. Technol. INTECH 2013, pp. 281–286, 2013.

[25] N. Rachmat and Samsuryadi, "Performance analysis of 256-bit aes encryption algorithm on android smartphone," J. Phys. Conf. Ser., vol. 1196, no. 1, 2019.

[26] N. Su, Y. Zhang, and M. Li, "Research on data encryption standard based on AES algorithm in internet of things environment," Proc. 2019 IEEE 3rd Inf. Technol. Networking, Electron. Autom. Control Conf. ITNEC 2019, no. Itnec, pp. 2071–2075, 2019.

[27] H. K. Kim and M. H. Sunwoo, "Low Power AES Using 8-Bit and 32-Bit Datapath Optimization for Small Internet-of-Things (IoT)," J. Signal Process. Syst., vol. 91, no. 11–12, pp. 1283–1289, 2019.

[28] V. K. Sarker, T. N. Gia, H. Tenhunen, and T. Westerlund, "Lightweight Security Algorithms for Resource-constrained IoT-based Sensor Nodes," IEEE Int. Conf. Commun., vol. 2020-June, 2020.

[29] I. I. Conference, "OpBench: A CPU performance benchmark for ethereum smart contract operation code," Proc. - 2019 2nd IEEE Int. Conf. Blockchain, Blockchain 2019, pp. 274–281, 2019.

[30] S. S. Brimzhanova, S. K. Atanov, M. Khuralay, K. S. Kobelekov, and L. G. Gagarina, "Cross-platform compilation of programming language Golang for Raspberry Pi," ACM Int. Conf. Proceeding Ser., vol. Article 10, pp. 1–5, 2019.

[31] C. Wang et al., "Go-Clone: Graph-embedding based clone detector for Golang," ISSTA 2019 - Proc. 28th ACM SIGSOFT Int. Symp. Softw. Test. Anal., pp. 378–381, 2019.

An Iterative, Self-Assessing Entity Resolution System: First Steps toward a Data Washing Machine

John R.Talburt[1], Awaad K. Al Sarkhi[2]
College of Science, Technology, Engineering, and
Mathematics, The University of Arkansas at Little Rock
Little Rock, USA

Leon Claassens[4]
PiLog Group
Centurion,
South Africa

Daniel Pullen[3]
Noetic Partners,
New York, NY

Richard Wang[5]
Sloan Management School, MIT
Boston, MA

***Abstract*—Data curation is the process of acquiring multiple sources of data, assessing and improving data quality, standardizing, and integrating the data into a usable information product, and eventually disposing of the data. The research describes the building of a proof-of-concept for an unsupervised data curation process addressing a basic form of data cleansing in the form of identifying redundant records through entity resolution and spelling corrections. The novelty of the approach is to use ER as the first step using an unsupervised blocking and stop word scheme based on token frequency. A scoring matrix is used for linking unstandardized references, and an unsupervised process for evaluating linking results based on cluster entropy. The ER process is iterative, and in each iteration, the match threshold is increased. The prototype was tested on 18 fully-annotated test samples of primarily synthetic person data varied in two different ways, good data quality versus poor data quality, and a single record layout versus two different record layouts. In samples with good data quality and using both single and mixed layouts, the final clusters had an average F-measure of 0.91, precision of 0.96, and recall of 0.87 outcomes comparable to results from a supervised ER process. In samples with poor data quality whether mixed or single layout, the average F-measure was 0.78, precision 0.74, and recall 0.83 showing that data quality assessment and improvement is still a critical component of successful data curation. The results demonstrate the feasibility of building an unsupervised ER engine to support data integration for good quality references while avoiding the time and effort to standardize reference sources to a common layout, design, and test matching rules, design blocking keys, or test blocking alignment. Also, the paper proposes how unsupervised data quality improvement processes could also be incorporated into the design allowing the model to address an even broader range of data curation applications.**

Keywords*—*Unsupervised entity resolution; data curation; frequency blocking; entropy regulated; data washing machine

I. INTRODUCTION

As organizations ingest and process larger amounts of data, the time and effort it takes to prepare and integrate data into useful products are also increasing, and many researchers are working to alleviate this bottleneck using several different approaches [1], [2], [3]. The root cause of the time delay is human supervision of the curation steps including data quality analysis, data cleansing and standardization, entity resolution (ER), and data integration [4]. The goal of ER is to link two references if, and only if, the references are equivalent [5], [6]. The problem is only exacerbated by Big Data [7], [8]. Because of the time delay between receiving data and its availability for use, data analysts often face the choice of waiting for the preparation to be complete, or to by-pass the curation process and engage in their attempts at data preparation which may or may not follow the best practices.

Many organizations are beginning to recognize this time and effort gap between data ingestion and final information product, and are moving to remedy this situation by increasing the level of automation in data curation processes [9]. These organizations along with software vendors and university researchers are trying to understand how to apply the same AI and ML techniques used for the data analytics at the end to the automation to the preceding data preparation processes [10], [11]. While many of these employ AI and ML [12], [3], [13], they still largely rely on some level of standardization in the source data. The ultimate goal is to develop systems for unsupervised data curation (UDC) which are metadata agnostic and can directly ingest and process raw data. The objective of UDC is to develop methods and techniques to process data at scale and successfully produce information products without manual intervention. Key components of the data curation process and prime targets for automation are the largely manual processes of data quality analysis, building transformation for data cleansing and standardization, and developing and testing rules for entity resolution and data integration (fusion).

UDC has been likened to a “data washing machine” [14]. When using a household washing machine for laundry, the user first loads the dirty laundry, and detergent then selects the cycles. The washing machine automatically executes the cycles, and in the end, produces clean laundry. Similarly, the user of the data washing machine loads dirty data with appropriate reference data, then selects the data cycles (control parameters). The data washing machine then executes the cycles to produce clean data (an information product) appropriate for use in a particular application.

The focus of this research is to describe a proof-of-concept (POC) prototype to serve as both a starting point and a foundation upon which a more complete UDC can be built [15]. The primary goal is to develop unsupervised methods and techniques for both data cleaning and data integration (ER) capable of operating at scale. The current code for the POC described in this paper can be found at https://bitbucket. org/Awaad_Al_Sarkhi/dwm-datawashingmachine/src/master/

II. A Proof-of-Concept (POC) for Unsupervised Data Curation (UDC)

The purpose of the POC is to demonstrate the feasibility of cleaning and integrating entity references in an automated fashion for certain types of data and certain phases of the curation process. The primary use case addressed by the POC is “multiple sources of the same information” as described in [16] as one of ten root causes of data quality problems. The novelty of the POC is it attempts to perform unsupervised entity resolution (ER) first rather than data cleaning, the opposite of most supervised processes. The objective of the POC is to minimize human intervention to analyze and transform the data and still obtain usable results as measured by the accuracy of clustering, i.e. a working data washing machine for data deduplication.

The POC for the data washing machine was written in Python and Java and uses frequency-based blocking, a multi-token scoring matrix as its ER matching process, and entropy-based quality evaluation of clustering [17] , [18] , [19], [20] The assumptions of the POC are

- The input to the process is a text file in a comma-separated values (CSV) format.
- Each text line is a reference to the same type of entity such as person entities (patients, customers, students), business entities, or materials (product listings, machine parts).
- The references are not assumed to be standardized with a uniform metadata tagging. No metadata is used in the POC process. Any metadata in the form of a header record is discarded.
- The first string value in each text line is a unique reference identifier.

To facilitate experimentation with various unsupervised techniques, the POC was developed as a series of sub-processes or phases to facilitate experimentation. Currently, phases have been implemented, and the fourth phase for token correction is under development. The organization of this paper is as follows:

- Phase I: Punctuation removal, upper casing, and tokenization.
- Phase II: Global standardization (replacement) of non-numeric tokens at the file level.
- Phase III: Removal of stop words, blocking, and clustering of equivalent references (entity resolution).

A. Phase I - Tokenization

The first Phase reads each reference as a line of text and performs a series of operations. The first is to separate the reference identifier, convert all letters to uppercase, and replace the field delimiters (typically a comma) with a blank character. Next, all non-word characters (\W) are replaced. For experimentation, two methods of replacement for non-word characters were tried. In the first method called "Compress," the non-word characters are replaced by a null character. For example, if a field has the value "123-456", then after replacing the hyphen character with a null character is becomes the single string "123456". In the second method called "Splitter," each non-word character is replaced by a blank character. The same example "123-456" becomes two strings (tokens), "123" and "345".

The motivation for the Compress method was to transform characteristic values with punctuation such as telephone numbers and dates into a single string. Interestingly, for the data used for the initial validation of the POC, the Splitter method generally gave better results than the Compress method.

In addition to non-word character replacement, upper casing, and tokenization, the first Phase also has an option to de-duplicate tokens. If the duplicate token option is employed, any duplicates of tokens within the same reference are removed, otherwise, duplicates are left in the reference. In the end, the cleaned tokens from each reference are reassembled into a blank delimited string and written to the tokenized reference file.

B. Phase II – Global Token Replacement

Phase II attempts an unsupervised correction of misspelled tokens based on the token frequency and string similarity. The replacement uses the assumption, if a high frequency, the non-numeric token is very similar to a low-frequency, non-numeric token, the low-frequency token is likely to be a misspelling of the high-frequency token and can be replaced by the high-frequency token. The validity of this assumption is dependent upon several factors. These include, what is a high frequency, what is a low frequency, and what is very similar.

The process is controlled by four parameters:

- MinFreqStdToken – The minimum frequency of a token that can be used to replace another token, i.e. can function as a "standard" token.
- MinLenStdToken – The minimum string length of a standard token.
- MaxFreqErrToken – The maximum frequency of a token that can be replaced by a standard token, i.e. can be treated as an “error” token.
- MaxStringDist – The maximum string (character) distance between a standard token and an error token before the error token can be replaced (usually 1 as measured by Levenshtein edit distance).

The replacement table has a one-to-many relationship between standard tokens and error tokens. One standard token could replace many different error tokens, but an error token can only be replaced by one standard token. Some actual examples of rows from the Replacement Table for Sample S8 are shown in Table I where

- MinFreqStdToken =10
- MaxFreqErrToken = 3
- MinLenStdToken = 4
- MaxStringDiff = 1

Table I shows some examples of token replacements generated in Phase II. It is important to note the token changes made in Phase II are not permanent changes to the source data. The token changes in Phase II are intended to improve the cluster (ER) results in Phase III.

TABLE I. EXAMPLE ROWS FROM REPLACEMENT

	Std Token	Freq	Error Token	Freq
1	APT	82	APTZ	1
2	APT	82	APLT	1
3	APT	82	APTR	1
4	CALIFORNIA	58	CALFORNIA	3
5	CALIFORNIA	58	CALIFORANIA	1
6	TEXAS	48	TEAS	3
7	TEXAS	48	TEXAYS	1
8	APARTMENT	32	PARTMENT	2
9	APARTMENT	32	APARTMENTS	1

Research is continuing on the development of Phase IV to make more accurate token corrections (standardization) at the cluster level. If it can be demonstrated the clusters produced by Phase III are reasonably accurate, then the criteria for identifying misspellings described for Phase II can be more aggressive when applied at the cluster level than at the file level. For example, while the replacements shown in Table I at the file level risks overwriting valid tokens, the same replacement is more likely to be valid within a cluster of references believed to be for the same person. Changes at the cluster level could also be applied to numeric tokens. For example, if five out of six references in a cluster have the token "413", and the sixth reference has “431”, and all six instances are preceded and followed by the same token, then it is not unreasonable to assume “431" is a mistyped version of “413”.

C. Phase III – Clustering (ER)

The purpose of Phase III is to cluster records for the same entity in support of data deduplication and data integration. This phase is more complex than Phases II and III and involves iterating over the tokenized source records coming out of Phase II. The clustering phase is a series of 13 processes labeled P1 through P13.

1) Process P1: Tokenize and compute token frequencies: Because the references have already been tokenized in Phase I, the re-tokenization here is simply a matter of separating the reference identifier and splitting the remaining substring by blank (white) space. While computing token frequencies is redundant with the same process in Phase II, for experimental purposes this was done to make Phase II an optional process allowing the evaluation of data integration results with and without token replacement.

2) Process P2: Tokenizing references and appending blocking tokens: Process P2 is the start of an iterative process on the “reprocess file”. Initially, the reprocess file is a copy of the original input file from which the frequency dictionary was created in Process P1. However, as the POC progresses, the reprocess file becomes a smaller and smaller subset of the original input source until there are there no more references to the process ending the iterations.

Process P2 repeats the tokenization process described in Process P1 in which each reference is split into a list of tokens. However, Process P2 has access to the token frequency dictionary previously build in P1. Process P2 has two primary functions:

- To rebuild each input reference as a string of blank-separated tokens, omitting all tokens found to have a frequency above the stop word frequency threshold (σ) creating “skinny references.”
- To output a copy of the skinny reference for each blocking token found in the reference.

Again, a blocking token is simply any token with a frequency below the blocking frequency threshold β. This means the output from P2 will have more records than the input assuming almost all references have at least one blocking token, and many have more than one.

Example: Suppose an input reference has the form

R13, John Doe, Oak St, Anyville AL, 793-1234

The tokenization of this reference would produce 9 tokens “R13”, “JOHN”, “DOE”, “OAK”, “ST”, “ANYVILLE”, “AL”, “793”, and “1234” (using Splitter tokenization). Also, suppose the tokens “JOHN”. “DOE”, and “OAK” have a frequency below β, and the tokens “AL”, “ST”, and “793” have a frequency above σ. Then P2 will generate three outputs.

R13: JOHN: JOHN DOE OAK ANYVILLE 1234

R13: DOE: JOHN DOE OAK ANYVILLE 1234

R13: OAK: JOHN DOE OAK ANYVILLE 1234

Because the input reference R13 contains three blocking tokens, P2 will output three skinny references, one for each blocking token. To simplify parsing, the output reference is divided into three segments using the colon (:) character. The first segment is the reference identifier, the second the blocking token, and the third the body of the reference.

3) Process P3: Sorting by blocking tokens to create blocks: The purpose of Process P3 is to sort the output of the reference from process P2 into ascending order by blocking token (Segment 2 of the rebuilt references). Each sequence of consecutive references with the same blocking token will form a block for input to the ER process for record linking.

4) Process P4: Iterate blocks: Process P4 is the start of an iterative process (P5) to be performed on each block. P4's primary function is to detect the sequences of consecutive records having the same blocking token, then pass this block of references on to P5.

5) Process P5: Link reference pairs in blocks: In Process P5, each block undergoes a process to generate pairs of linked references. The technique implemented in the POC is to use a multi-token comparator. Every pair of references in the block is compared. For a block of N references, there will Nx(N-1)/2 pairs.

Any pairwise matching process can be inserted at this point including machine learning (ML) algorithms for linking. Because the entity references are text, this approach usually requires an additional process to convert references from the text to numeric vectors, a process called text embedding. Some results from using the DBScan clustering algorithm with doc2vec text embedding are shown in this paper (Table III).

Most of the work described here used the scoring matrix. In this case, a variation of the Monge-Elkan method [21] for comparing multi-token values, but with the removal of stop words. When the scoring matrix processes a pair of references, each reference is first transformed into a list of tokens (words), then the stop word tokens are removed from the list. The remaining tokens from the first reference are used to label the rows of the matrix, and the remaining tokens from the second string label the columns of the matrix. The cell value of the matrix is a normalized similarity measure, i.e., a value in the interval [0,1], between the two tokens. In the POC, the normalized Damerau-Levenshtein Edit Distance (nLED) function was used.

To illustrate the operation of the scoring matrix, consider the following two references:

A045, Smith, John, Apt 21, 345 Oak St, Anytown, NY

B167, Jon Smith, 345 Oak Street #21, Anytown, NY

Furthermore, suppose the threshold for the comparator has been set to 0.80, and the list of stop words contains the token "NY." The resulting token matrix would then appear. The process begins by finding the largest similarity value in the matrix. This value is the initial value of a total running value. After the largest similarity value is used to initialize the total value, all of the values in the same row and column are removed (set to zero). In the next iteration, the largest similarity value from the remaining values in the matrix is identified and added to the overall total.

Again, all of the nLED values in the same row and column as the largest value are removed. The process continues in subsequent iterations until all of the similarity values have been removed from the matrix. In Fig. 1, the cells with underlined and bold font are the surviving similarity scores from this process. After the last iteration, the running total is divided by the number of iterations. If the calculated average value is greater than or equal to a threshold value provided by the user, then references are linked. At the end of the algorithm, the final matrix score for a pair of references in Fig. 1 is 0.83.

	JON	SMITH	345	OAK	STREET	21	ANYTONW
SMITH		**1.00**			0.17		0.14
JOHN	**0.75**			0.25			0.14
APT		0.20			0.17		0.29
21						**1.00**	
345			**1.00**				
OAK				**1.00**			0.14
ST		0.40			**0.33**		0.14
ANYTOWN	0.29	0.14		0.14			**0.71**

Fig. 1. Example Scoring Matrix (Zero Similarity Values Omitted).

6) Process P6: Linked pair generation: The purpose of Process P6 is to form the graph edges between pairs of references in the same cluster. Because the clusters are all formed from references in the single token block, they only represent the connections found between references sharing the token forming the block.

7) Process P7: Post-resolution transitive closure: Unlike traditional match key blocking, frequency-based blocking does not produce a true partition of the input references where each input reference is in one, and only one, block. In frequency-based blocking, each reference is replicated by the number of blocking tokens it contains as in the example for Process P3. To create the final set of clusters, in which each reference occurs in one, and only one, cluster, the clusters created from the blocks must be merged and undergo a transitive closure process.

The POC implements a very efficient sorting closure process described by Kolb et al [22]. While the sorting transitive closure is implemented in the POC as an in-memory, Java application, the algorithm is a highly-scalable, map/reduce process for execution in the Hadoop Distributed File System (HDFS) environment.

8) Process P8: Iterate clusters: Process P8 transforms the transitive closure output into clusters of linked references. Because the output of the sorting closure process is already in sorted order by cluster identifiers, the clusters are simply groups of consecutive references with the same cluster identifier.

9) Process P9: Entropy Calculation: Process P9 uses a variation of the Shannon entropy calculation [23] to assess the level of organization in each cluster of two or more references. The formula for the calculation of the entropy of a cluster is

$$\mathrm{E} = -\sum_{j=1}^{N} p\left(t_j\right) \cdot \log_2 p\left(t_j\right) \qquad (1)$$

Where t_j is the j-th vertical token group in the cluster, and $p(t_j)$ is the probability of t_j.

For the POC, a vertical token group is defined to be the same token counted only once in each reference of a cluster. Thinking of the cluster as a matrix where the references are the rows and the columns are the tokens, then a vertical token group is a vertical grouping of the same token across different references. However, each token is counted only once in each reference. This means the maximum size of a vertical token group is equal to the number of references in the cluster. The probability of a vertical token group is the size of the token group divided by the number of references in the cluster. For example, consider the following cluster of 3 references.

R1: JOHN GRANT 123 GRANT ST

R2: MARY GRANT 21 OAK STREET

R3: MARY GRANT 21 OAK ST

The first vertical token group is for the token "JOHN" which only occurs once in R1 forming a vertical token group of size 1 with a group probability of 1/3. The second vertical token group is for "GRANT" which has 3 tokens, one token each from R1, R2, and R3 giving this group a probability of 1.0 (3/3). The second "GRANT" in R1 is not part of this token group because each token is only counted once in each reference. The token group for "123" has a probability of 1/3, the second "GRANT" group has a probability of 1/3, and the "ST" group a probability of 2/3.

After exhausting all of the tokens in R1, there are still four uncounted tokens in R2 forming the "MARY" group with probability 2/3, "21" group probability 2/3, the "OAK" group probability 2/3, and the "STREET" group probability 1/3. Finally, there are no remaining uncounted tokens in R3. In total, there are 9 vertical token groups in the example cluster. The total entropy of the cluster is calculated from Formula (2) by:

$$E = -(\frac{1}{3}\cdot log\left(\frac{1}{3}\right) + 1\cdot log(1) + \frac{1}{3}\cdot log\left(\frac{1}{3}\right) + \frac{1}{3}\cdot log\left(\frac{1}{3}\right) + \frac{2}{3}\cdot log\left(\frac{2}{3}\right) + \frac{2}{3}\cdot log\left(\frac{2}{3}\right) + \frac{2}{3}\cdot log\left(\frac{2}{3}\right) + \frac{2}{3}\cdot log\left(\frac{2}{3}\right) + \frac{1}{3}\cdot log\left(\frac{1}{3}\right)) \; E = 3.67318$$

Entropy is a measure of the organization of a cluster in terms of having similar tokens [24]. The entropy of a cluster decreases as references in a cluster have more and more similar tokens. By this measure, a cluster will have an entropy of 0 if, and only if, all of the references have the same tokens.

10)Process P10: Assessment of clusters based on entropy: In Process P10, the entropy of each cluster as calculated in Process P9 is assessed against the user-defined entropy threshold ε. If a cluster has an entropy less than ε, it is judged as an acceptable cluster, and the reference identifier and cluster identifiers from the clusters are written to the Saved Clusters output file. Otherwise, the cluster identifiers are discarded, and the references are written to the Reprocess file. References written to the Reprocess file will go through the entire blocking and ER process again but at a higher match threshold. By definition, all clusters of size one (singleton clusters) have an entropy of 0 and are written directly the Saved Clusters file.

For each cycle of the POC, the size of the Saved Clusters file increases while the size of the Reprocess file decreases. The Reprocess file will eventually become empty as the match threshold μ approaches 1.0. At very high match thresholds, the references in a block can only form clusters if they are highly similar and generate clusters of very-low entropy, otherwise, they break down into singleton clusters. In either case, they will eventually pass to the Saved Clusters file and the Reprocess file will be empty. An example of this process is shown in Table I. The statistics are produced as part of the statistics report when running a sample. In this case, the statistics are for Sample S4 of 1,912 references. As shown in Table IV, the parameters for this run were β=12, σ=22, ε=4.2, and the starting value of μ=0.5.

The volume of work continually decreases with each iteration. Note that some references written to the reprocess file will not be used in the next iteration. This is because, at the beginning of the next iteration, the reprocess file is re-blocked and re-clustered. During the clustering process, reference-to-reference links are only produced for references linked to at least one other reference. For example, 27 references were written to the reprocess file at the end of the μ=0.8 iterations, but only 14 of these references survived to form 6 clusters of two or more references when the match threshold μ was increased to 0.90.

11)Process P11: Reprocess decision: As described in Process P10, at some point the Reprocess file will be empty. When this happens, the reprocessing cycle stops, and the final join (Process P13) is performed.

12)Process P12: Increasing Match Threshold: If there are references to be reprocessed, then the match threshold is increased before the reprocess is started. Increasing the match threshold will require references to be more similar before they are linked into the same cluster. In all of the results reported here, the increment value was 0.1 (10%).

13)Process P13: Final join to original source: Although no further iterations are necessary when the Reprocess file is empty, there are still two tasks to complete. The first task is to ensure every reference in the source is represented in the final set of clusters. Some references in the source may not be transferred to the Saved Clusters file. Depending upon the value of blocking frequency threshold β, some references may not contain blocking tokens and are not output from Process P2.

The second task is to append the final cluster identifier to each reference in the source. The goal is to create a Final output comprising every reference in the source along with its proper cluster identifier. Both of these tasks can be completed by performing an outer join by reference identifiers between the original Reference Source file and the Saved Clusters file.

D. Cluster Cleaning

While this process has not been implemented in the POC, work is currently underway to develop unsupervised techniques for cleaning and standardizing tokens within the same cluster. The current approach is very much the same as the Global Token Replacement described in the Second Phase. However, replacements can be more aggressive at the cluster-level versus the file level. Across an entire reference file, there could easily be an entire sequence of house numbers, such as 123, 124, 125, and so on. For this reason, numeric tokens are specifically excluded from replacement globally. However, at the cluster level, it is much more probable that if 5 of 6 references have the token 123 and the sixth reference has 124, the replacement of 124 by 123 would be a correction.

III. POC Test Samples and Results

To test the POC, 18 samples were taken from four fully annotated reference sources. Aside from having equivalent references to match, the samples also exhibited combinations of two other characteristics – high data quality (DQ=Good) versus low data quality (DQ=Poor) and uniform record layout (Mixed=No) versus mixed (Mixed=Yes) record layout. In all cases, the Splitter Tokenization Method with token de-duplication was used in Phase I. To gauge the effect of Phase II (Global Token Replacement) each sample was run with, and without, the global token replacement. In the cases where token replacement was run, the settings described in Section on the Second Phase the parameters were fixed at

- MinFreqStdToken =5
- MaxFreqErrToken =3
- MinLenStdToken = 5
- MaxStringDiff = 1

To establish a baseline, all samples were run with 0.50 as the initial value of μ and 0.10 as the increment value forμ. The initial values for β and σ were set using the linear regression prediction formulas (3) and (4) for the non-iterative model [25]. However, the actual values for β, σ, and ε were set manually by observing the correlation between the F-measure of each cluster and the computed entropy as logged by the system (Table IV). Then exploring a range of values around these estimates using a grid search automated with a robotic Python process to run each range of settings and collect the precision, recall, and F-measure results. The results for all 18 samples using the best parameter settings are given in Table IV.

A. *Samples with Good Data Quality*

Stratified samples S1, S2, S4, S5, S7, S13, S14, and S15 were drawn from a corpus of approximately 800K references created using the R-package “generator”, and degraded with data quality errors using the R-package “relErrorGeneratoR” from GitHub.com. While some reference-level errors such as misspelling, truncation, mixed formatting, and missing values were injected into the data during generation, the individual references in the 800K corpus are of relatively high quality.

The majority of the data quality errors introduced into the 800K corpus were data redundancy (duplicate record) errors to make the corpus more useful for entity resolution research. Shown here are two references from Sample S4 with Record Layout A. The only variations between the two references are the name truncation (initial) and different formats for telephone numbers and identification numbers.

A926344: ANDREW, AARON, STEPHEN, 2475 SPICEWOOD DR, WINSTON SALEM, NC, 27106, 601-70-6106, (159)-928-5341

A930444: A, AARON, STEPHEN, 2475 SPICEWOOD DR, WINSTON SALEM, NC, 27106, 601706106, (159)9285341

Sample S6 was produced by the GeCo synthetic data generator [26], and Sample S3 is a file of 866 references to restaurants (businesses) from two public sources, Zagat’s and Fodor’s restaurant guides. The references contain restaurant names, addresses, city, phone, and type of cuisine. The file has been manually annotated and is known to have 112 pairs of equivalent references [27]. Examples of references from S3 are shown here.

A001: Arnie Morton's of Chicago 435 S. La Cienega Blvd. Los Angeles 310-246-1501 Steakhouses

A002: Arnie Morton's of Chicago 435 S. La Cienega Blvd. Los Angeles 310/246-1501 American

B. *Low Data Quality Samples*

Samples S9, S10, S11, S12, S16, S17, and S18 were taken from the SOG (Synthetic Occupancy Generator) project [28]. The SOG corpus has approximately 270K references with three different record layouts A, B, and C. The SOG corpus has a much higher level of data quality errors than the 800K corpus. Most records exhibit at least one error such as missing value, misspelling, truncation, inconsistent formatting, nicknames, and name and address changes. Shown here are three equivalent references from Sample S8 exhibiting a number of these data quality issues.

A960175,lucia,r,oster,t20672,southwood,oaks,dr,porter,,tx, 77365,,,10896980,,

A966807,lucia,r,wilson,12006,MOUNTAIN,RIDGE,RD,HOUSTON,,TEXAS,77043,PO,BOX,280034,
houston,,tx,77228,10896980,1917

A971069,LUCIA,R,WILSON,20672,SOUTHWOOD,OAKS,DR,PORTE,,TEXAS,77365,,,001-89-6980,,

C. *Mixed Layout Samples*

In addition to variations in quality, Sample S7 and Samples S10 through S18 were selected with mixed (heterogeneous) record layouts. For example, in Sample S7 about half of the references were in Record Layout A and the other half in Record Layout B. The two layouts used a different order for names and have different identity attributes, e.g. social security number in Layout A and date-of-birth in Layout B. An example of a pair of references from S7 is shown here.

A944353,VICTOR,AGWU,KINGSLEY,1608 W NORTHWEST BLVD # O,WINSTON SALEM,NC,27104.0,730-69-2869

B867674,VICTOR K AGWU,1608,W NORTHWEST BLVD # O,WINSTON SALEM,NC,27104,12/14/37,

For samples selected from the SOG corpus, the difference in layouts was much greater. Here is an example of a pair of references from Sample S10.

A993286,chavez,4149 WALSH LN,GRAQND PRAIRIE, TEXAS 75052,OFFICE BOX 54331,grand prairie, tx 75054,10525947,,,

A994281,barbie chavze,11R881 GULF POINTE DR APT E38,HOUSTON, TX 77089,,,,,,, B904140,Barby ,CHAVEZ ,11881,GULF POINTE DR ,Apt E38,HOUSTON,TX,77089,(713)165.7474,,

As noted, the values for β, σ, ε, and the starting value of μ were found by a grid search. Prior research in using the scoring matrix for ER on samples from these same corpora [25] [29], [30], [31] provided some guidance of the best values for the blocking threshold frequency threshold β and the stop word frequency threshold σ based on the size of the sample and the standard deviation of its token frequency distribution. These formulas are

$$\beta = (7.8864) + (0.0023)*Std\_Dev + (0.0005)*Size \qquad (3)$$

$$\sigma = (-83.3106) + (2.9647)*Std\_Dev + (0.0037)*Size \qquad (4)$$

However, the previous research did not involve the entropy-based, self-evaluation, or iteration with incrementally increasing match thresholds use in this research, thus it did not provide any guidance about the best setting for the entropy threshold ε. Instead, the estimated value for ε was found by observing the entropy measure of each cluster and comparing it to the actual F-measure of the cluster. This was possible because all of the test samples were fully annotated. The F-measure assessment of each cluster was an augmentation to Processes P5. As the entropy of each cluster is calculated, the cluster was also sent to an ER metrics program to determine the actual F-measure of the cluster as compared to the annotated truth set.

The entropy and the actual F-measure of each cluster were captured in a Cluster Analysis text file. Table II shows a segment of the report produced when running Sample S2. The table shows the results of three iterations. Row 1 of Table II shows the last cluster produced by the initial value μ at 0.5, Rows 2 through 6 show the entire second reprocess iteration of four clusters where the value μ was 0.6. Row 7 is the first cluster of the last iteration where μ was 0.7. As each cluster is formed, its entropy is calculated as shown in the column labeled "Entropy." If the cluster's entropy is above the value of ε (set at 4.3 for this run) the cluster is judged to be "bad" and is written to the reprocess file for re-linking at the next higher value of the match threshold μ.

On the other hand, if the entropy is less than or equal to ε, it is written to the "good" file as a final cluster. Table II shows for Rows 4 and 5 these were correct decisions. In both cases, the F-Measure was less than 1.0 when the entropy was above 4.3. However, Row 2 is an exception. Even though the entropy of 9.0446 is above 4.3, the cluster had an F-Measure of 1.0 and was correctly linked. However, because the entropy was above the threshold, the references were put back for reprocessing in the third iteration. In the end, the F-Measure for S2 at the end of the process was 0.8842 as shown in Table II.

TABLE II. SEGMENT OF ENTROPY VS. F-MEASURE REPORT FOR S2

Row	μ	ε	Size	Entropy	F-Meas	Precision	Recall
…							
1	0.5	4.3	2	0.00	1.0	1.0	1.0
2	0.6	4.3	3	9.04	1.0	1.0	1.0
3	0.6	4.3	2	1.00	1.0	1.0	1.0
4	0.6	4.3	3	5.53	0.5	0.33	1.0
5	0.6	4.3	4	5.50	0.5	0.33	1.0
6	0.7	4.3	3	9.04	1.0	1.0	1.0
7	0.7	4.3	2	2.00	1.0	1.0	1.0
…							

D. Example Results using Machine Learning for P5

In this example, Sample S4 was processed using DBScan (Density-Based Spatial Clustering of Applications with Noise) [32] as the ML clustering algorithm and using the doc2vec [33] word embedding algorithm to create the numeric vectors as input for DBScan. As implied by its name, the doc2vec algorithm converts an entire document into a vector. For the POC, each reference was considered a document so there is a one-to-one correspondence between each input reference and each vector clustered by DBScan.

The doc2vec algorithm was applied to each block using the following parameters.

- vector size = 5
- min count = 2
- epoch = 20
- alpha = 0.25
- min_alpha=0.00025

DBScan algorithm was imported from the Python 3.7 library learn. cluster. This version has two control parameters "eps" and "min_samples". The eps parameter controls the neighborhood reach (proximity) of vectors to be in the same cluster, and the min_samples parameter defines the minimum size of "core samples", i.e. the minimum number of vectors within eps distance of each other. The results from using this configuration are shown in Table III.

TABLE III. ER RESULTS USING DOC2VEC FOLLOWED BY DBSCAN

Sample	Size	eps	min_ samples	Results		
				Precision	Recall	F-Measure
S1	50	0.6	0.9	0.8519	1.0000	0.9200
S2	100	0.4	1	0.9070	1.0000	0.9513
S4	1,912	0.01	1	0.9555	0.9111	0.9328

TABLE IV. SHOWS THE RESULTS FROM THE POC

ID	Size	Token	DQ	Mix	Mu	Mu+	Beta	Sigma	Epsilon	Precision	Recall	F-Measure
S1	50	Split	Good	No	0.50	0.10	6	7	4.9	1.00	0.963	0.9811
S2	100	Split	Good	No	0.50	0.10	6	7	4.7	1.00	0.8958	0.9451
S3	868	Split	Good	No	0.50	0.10	9	95	4.0	0.8889	0.9286	0.9083
S4	1,912	Split	Good	No	0.50	0.10	12	22	3.1	0.9854	0.8869	0.9335
S5	3,004	Split	Good	No	0.50	0.10	12	53	3.0	0.9911	0.8729	0.9282
S6	19,998	Split	Good	No	0.50	0.10	35	403	15.1	0.9457	0.9737	0.9595
S7	3,000	Split	Good	Yes	0.50	0.10	14	24	3.0	0.9464	0.8665	0.9047
S8	1,000	Split	Poor	No	0.60	0.05	14	145	35	0.7880	0.8881	0.8350
S8	1,000	Comp	Poor	No	0.60	0.05	14	144	33	0.7877	0.8827	0.8324
S9	1,000	Split	Poor	No	0.50	0.05	15	135	28	0.6824	0.8453	0.7552
S9	1,000	Comp	Poor	No	0.62	0.02	23	150	28	0.7806	0.8116	0.7958
S10	2,000	Split	Poor	Yes	0.50	0.05	31	280	31.5	0.7235	0.8348	0.7752
S10	2,000	Comp	Poor	Yes	0.62	0.02	32	280	34	0.7455	0.9041	0.8172
S11	4,000	Split	Poor	Yes	0.55	0.02	38	280	31.5	0.7010	0.8386	0.7636
S11	4,000	Comp	Poor	Yes	0.62	0.02	43	258	26	0.7571	0.7764	0.7666
S12	6,000	Split	Poor	Yes	0.50	0.05	20	580	29	0.7699	0.7424	0.7560
S12	6,000	Comp	Poor	Yes	0.62	0.02	21	570	26.4	0.7825	0.7723	0.7774
S13	2,000	Split	Good	Yes	0.50	0.05	14	23	5.3	0.9478	0.8116	0.8745
S13	2,000	Comp	Good	Yes	0.01	0.02	14	110	2.4	0.9707	0.8003	0.8773
S14	5,000	Split	Good	Yes	0.48	0.02	24	118	5.2	0.9419	0.8464	0.8916
S14	5,000	Comp	Good	Yes	0.01	0.02	24	310	2.6	0.9579	0.8226	0.8851
S15	10,000	Split	Good	Yes	0.50	0.02	20	100	5	0.9543	0.8242	0.8845
S15	10,000	Comp	Good	Yes	0.1	0.03	27	108	2.8	0.9622	0.8247	0.8882
S16	2,000	Split	Poor	Yes	0.50	0.05	21	142	32.8	0.6908	0.8160	0.7666
S16	2,000	Comp	Poor	Yes	0.54	0.02	30	154	31.3	0.6500	0.9331	0.7663
S17	5,000	Split	Poor	Yes	0.50	0.05	35	480	32.8	0.6954	0.81578	0.7508
S17	5,000	Comp	Poor	Yes	0.60	0.02	26	466	26	0.7560	0.80716	0.7807
S18	10,000	Split	Poor	Yes	0.50	0.05	33	449	31.8	0.6995	0.7514	0.7245
S18	10,000	Comp	Poor	Yes	0.52	0.02	26	444	34	0.7511	0.7948	0.7724

IV. CONCLUSION AND FUTURE RESEARCH

The results are shown in Table IV suggest entropy can be an effective way to regulate an unsupervised clustering process. The POC using the scoring matrix performs extremely well when processing good quality references such as Samples S1 – S7 and S13 – S15. The average F-Measure for these samples was 0.9124 with an average precision of 0.9609.

The average F-Measure for the poor quality samples S8 – S12 and S16 – S18 was somewhat lower at with an average F-Measure of 0.7772 and precision of 0.7351. The results also indicate the POC is more sensitive to data quality issues than to mixed record formats. The good-quality, mixed-format Samples S7, S13, S14, and S15 had an average F-Measure of 0.8866 compared to an average F-Measure of 0.9426 for good-quality, single format samples.

For the good quality samples where the clustering precision was 96%, the hope is that applying a more comprehensive cleaning and standardization at the cluster level will be able to provide much better results. The goal for future research is, just as linking results can be continually improved through iterative reprocessing, the same reprocessing loop will also incorporate processes to continually improve the quality of the references, which in turn, would further improve the linking results. The POC described in this paper shows the unsupervised ER improvement part of this positive feedback loop is feasible. The next step will be to integrate additional unsupervised data quality improvement processes.

A. Industry Testing

As an experiment, a commercial company tested the POC (data washing machine) approach using a real-world dataset of 70,500 business names and address references with mixed record layouts. Because the dataset was not annotated, it was not possible to calculate the exact F-measure of the overall clustering results. However, the company did undertake an extensive manual review of the POC results in comparison to results from their standard process. The company determined the POC results to be as good as, and in many cases better than, the results from their standard process, but with the added advantage of avoiding the time and effort to analyze and prepare the data required by their standard process.

Besides, the company is experimenting with some variations of the original POC design described in this paper. In particular, they have been able to improve the clustering accuracy for their datasets by using a computed value for ε, the entropy threshold. In their approach, they consider five factors when assessing the entropy of each cluster. These are

- The match threshold μ used to form the cluster.
- The numbers of references in the cluster (size).
- The maximum number of tokens in any one reference in the cluster (maxT).
- The minimum number of tokens in any one reference in the cluster (minT).
- The average number of tokens for all references in the cluster (avgT).

In Process P10, instead of comparing the entropy of the cluster to a static value of ε as in the original POC, they compute a dynamic threshold based on the factors listed above. In particular, ε is computed as

$$\varepsilon = \log(1 + sizeC) + mu \cdot \{\log(avgT) + \log(1 + maxT - avgT) + \log(1 + maxT - minT)\} \quad (5)$$

In another change, they were able to improve the precision of the clustering by modifying the scoring matrix used to link references in Process P5. The comparator was modified to use a Boolean similarity of the match (1.0) and no-match (0.0) when comparing numeric tokens while still using the normalized Damerau-Levenshtein edit distance when comparing non-numeric tokens.

B. Predicting Parameters and Scalability

However, there are still two gaps that must be bridged to make the POC a practical solution for most real-world use cases. The first is a reliable method for setting the optimal values of the key parameters β, σ, and ε. In a research environment using fully annotated references, these values can be found by simply observing where the best results were obtained based on comparisons with the correct linking. When working with real data, this is not generally possible. A practical unsupervised ER system needs a way to predict these parameters for a given set of input references. Creating such predictive models is still research in progress.

The second consideration is scalability. The current POC is implemented in a combination of Python and Java, and as written, it is not very scalable. The blocking and the stop word removal process can be combined with the token counting process to avoid the need for storing an in-memory token frequency table.

The POC can be converted to an HDFS Map/Reduce process. The references can easily be tokenized in the mapping process which then reduces on the token. The reducer can then emit two kinds of key-value pairs for each token group. The first is (RefID, Token) where the token has a frequency below σ (not a stop word). The second is (Token, RefID) where the token has a frequency below β (a blocking token). Sorting and reducing the first pairs on the RefID as the key will create the skinny references of Process P2 while sorting and reducing the second pairs on Token as the key will create the blocks. The join of these two outputs on RefID will be the equivalent of creating and sorting the blocked file in Process P3. Next, the blocks can be mapped to distributed nodes for pairwise linking in parallel with the assurance no block will be larger than β. The outputs are the Process P6 Linked Pairs. The transitive closure of the pairs in Process P7 using the algorithm of Kolb et al [22] is already an efficient map/reduce process. Process P8 then becomes a map of the clusters to parallel processing work nodes performing the entropy calculation (Process P9) and triage of clusters (Process P10) into "good" and "bad" cluster outputs.

The POC described here was built uses the simplest of approaches which could no doubt be dramatically improved through additional research and experimentation including investigating different starting values forμ, and exploring its sensitivity to the increment value currently fixed at 0.1. Also, building prediction models for these parameters. Another is investigating whether the results are improved by modifying the values of β, σ, or ε for each reprocesses iteration, and if it does, how should they be modified to produce the best linking results.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Award No. OIA-1946391 and by the PiLog Group. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the PiLog Group.

REFERENCES

[1] A. Saeedi, E. Peukert and E. Rahm, "Incremental Multi-source Entity Resolution for Knowledge Graph Completion," German Federal Ministry of Education and Research, Leipzig, Germany, 2020.

[2] M. Stonebraker and I. F. Ilyas, "Data Integration: The Current Status and the Way Forward," Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, pp. 3-9, 2018.

[3] R. Yuji, H. Geon and E. W. Steven, "A Survey of Data Collection for Machine Learning: A Big Data – AI Integration Perspective," IEEE Transactions on Knowledge and Data Engineering, 2019.

[4] F. Azzalini, S. Jin, M. Renzi and L. Tanca, "Blocking Techniques for Entity Linkage: A Semantics-Based Approach," Data Science and Engineering, pp. 1-19, 2020.

[5] A. Alsarkhi and J. R. Talburt, "A method for implementing probabilistic entity resolution," International Journal of Advanced Computer Science and Applications, vol. 9, no. 11, pp. 7-15, 2018.

[6] P. G. Ipeirotis, V. S. Verykios and A. K. Elmagarmid, "Duplicate record detection: A survey," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 1, pp. 1-16, 2007.

[7] Y. Roh, G. Heo and S. E. Whang, "A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective," IEEE

Transactions on Knowledge and Data Engineering, vol. Early Access, 2019.

[8] V. Christophides, V. Efthymiou, T. Palpanas, G. Papadakis and K. Stefanidis, " An Overview of End-to-End Entity Resolution for Big Data," ACM Computing Surveys (CSUR), vol. 53, no. 6, pp. 1-42, 2020.

[9] P. Neoklis, R. Sudip, E. W. Steven and Z. Martin, "Data Lifecycle Challenges in Production Machine Learning," A Survey, ACM SIGMOD, vol. 47, no. 2, 2018.

[10] B. J. Dooley, "How Robotic Process Automation Eases Data Management, The Data Warehouse Institute (TDWI)," 18 June 2018. [Online]. Available: https://tdwi.org/articles/2018/06/18/diq-all-how-robotics-process-automation-eases-data-management.aspx.

[11] J. Adams, ", Automating Data Management and Governance through Machine Learning, The Data Administration Newsletter (TDAN)," 7 November 2018 . [Online]. Available: https://tdan.com/automating-data-management-and-governance-through-machine-learning/23972.

[12] R. Pita, L. Menezes and M. Barreto, "Applying Term Frequency-Based Indexing to Improve Scalability and Accuracy of Probabilistic Data Linkage," In LADaS@ VLDB, pp. 65-72, 2018.

[13] A. Moshyedi, T. Kramer, A. Gangopadhyay and S. Pal, "A combined semantic search and machine learning approach for address entity resolution," EasyChair, 2019.

[14] R. Y. Wang, J. R. Talburt and Y. W. Lee, "A framework for analysis of data washing machines," http://mitiq.mit.edu, Cambridge, MA, 2020.

[15] A. Jurek-Loughrey and P. Deepak, ""Semi-supervised and unsupervised approaches to record pairs classification in multi-source data linkage," In Linking and Mining Heterogeneous and Multi-view Data, pp. 55-78, 2019.

[16] Y. Lee, L. Pipino, J. Funk and R. Wang, Journey to Data Quality, MIT Press, 2006.

[17] U. Draisbach, P. Christen and F. Naumann, "Transforming pairwise duplicates to entity clusters for high-quality duplicate detection," Journal of Data and Information Quality (JDIQ), vol. 12, no. 1, pp. 1-30, 2019.

[18] P. Christen, "A survey of indexing techniques for scalable record linkage and deduplication," IEEE transactions on knowledge and data engineering, vol. 24, no. 9, pp. 1537-1555, 2011.

[19] A. Ardalan, A. Doan and A. Akella, "Smurf: Self-service string matching using random forests," Proceedings of the VLDB Endowment, vol. 12, no. 3, pp. 278-291, 2018.

[20] A. Mazeika, M. H. Böhlen, N. Koudas and D. Srivastava, "Estimating the selectivity of approximate string queries," ACM Transactions on Database Systems (TODS), vol. 32, no. 2, pp. 12-es., 2007.

[21] A. E. Monge and C. P. Elkan, "The Field Matching Problem: Algorithms and Applications," in KDD-96 Proceedings, 1996.

[22] L. Kolb, S. E. and E. Rahm, "ITerative Computation of Connected Graph Components with MapReduce," Datenbank-Spektrum, vol. 14, no. 2, 2014.

[23] C. E. Shannon, "A Note on the Concept of Entropy," Bell Systems Technical Journal, 1948.

[24] D. Lee, L. C. Zhang and J. K. Kim, "Maximum Entropy classification for record linkage," arXiv preprint arXiv:2009., p. 14797, 2020.

[25] A. Al-Sarkhi and J. R. Talburt, "Estimating the Parameters for Linking Unstandardized References with the Matrix Comparator," Journal of Information Technology Management, pp. 12-26, 2019.

[26] K. N. Tran, D. Vatsalan and P. Christen, "GeCo: an online personal data generator and corruptor," the 22nd ACM International Conference on Information & Knowledge Management, pp. 2473-2476, (2013, October).

[27] T. S, "Restaurant Benchmark Dataset," [Online]. Available: http://www.cs.utexas.edu/users/ml/riddle/data.html].

[28] J. R. Talburt, Y. Zhou and S. Y. Shivaiah, "SOG: A Synthetic Occupancy Generator to Support Entity Resolution Instruction and Research," MIT International Conference on Informationi Quality, pp. 91-105, 2009.

[29] A. Alsarkhi and J. R. Talburt, "Optimizing Inverted Index Blocking for the Matrix Comparator in Linking Unstandardized References," in ," in in Proceedings of the 2019 International Conference on Scientific Computing, Las Vegas, 2019.

[30] A. Al Sarkhi and J. Talburt, "An analysis of the effect of stop words on the performance of the matrix comparator for entity resolution.," The Journal of Computing Sciences in Colleges, vol. 34, no. 7, pp. 64-71, 2019.

[31] A. Al Sarkhi and J. Talburt, "A Scalable, Hybrid Entity Resolution Process for Unstandardized Entity References," The Journal of Computing Sciences in Colleges, pp. 19-29, 2020.

[32] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proceedings of Second International Conference on Knowledge Discovery and Data Mining, Portland, OR,, 1996.

[33] J. H. Lau and T. Baldwin, "An empirical evaluation of doc2vec with practical insights into document embedding generation," arXiv preprint arXiv:1607.05368, 2016.

A Model for Traffic Management based on Text Mining Techniques

Ahmed Ibrahim Naguib[1], Hala Abdel-Galil[2], Sayed AbdelGaber[3]
Faculty of Computers and Artificial Intelligence
Helwan University, Helwan
Egypt

Abstract—**It is very important for traffic management to be able to correctly recognize traffic trends from large historical traffic data, particularly the congestion pattern and road collisions. This can be used to reduce congestion, improve protection, and increase the accuracy of traffic forecasting. Choosing the correct and effective text mining methodology helps speed up and reduces the time and effort needed to retrieve valuable knowledge and information for future prediction and decision-making processes. Modeling collisions or accident risk has also been an important aspect of traffic management and road safety, as it helps recognize problems and causes that contribute to a higher risk of accidents, promotes treatment delivery, and reduces crashes to save more lives and avoid road congestion. Therefore, this work-study proposed a model that relies on the different text mining methodology to determine clearly what circumstances affect and who is involved more in an accident. Using different classification and machine learning techniques applied to get the optimum classifiers used in this model. The experimental results on real-world datasets demonstrate that the proposed models outperform Prayag Tiwari's Research Work related to the Leeds UK Dataset.**

Keywords—Classification; machine learning; text mining; traffic management

I. Introduction

Traffic management is a major problem which almost daily affects us. Usage of technologies such as the Internet of Things (IoT) and image processing will result in a smooth traffic management system. The main cause of traffic congestion is the lack of an appropriate mechanism for prioritizing traffic. The IoT is an infrastructure network. There are switches, sensors, actuators, and circuits in the embedded systems. With software and connectivity locally or over the internet helps in the transfer of data which can be provided by ThingSpeak API that can get the data in CSV text format [1]. As outlined in different IoT applications, the deployment of incident notification systems is one of the most common and important technologies in the smart transport field [2].

The implementation of sensors and IoT devices in Smart Traffic Network helps collect user expectations and contextual details that may be in the form of travel time, weather conditions, or real-life driving patterns Once the congestion situation is expected, alternative congestion-free routes are proposed that can be propagated by text according to the desired criteria are suggested that can be propagated through text messages or e-mails to the users [3].

There are two forms of congestion: structural or incidental. Structural congestion arises when traffic demand is greater than available, while incidental congestion results from irregular circumstances such as accidents, bad weather, or road work that alters traffic flow [4]. The ability to predict the impact of an incident immediately after its occurrence is crucial to advanced traffic management and significantly improves the system's performance [5].

Present traffic flow management strategies may not be adequately successful to monitor the changing and continuing traffic as transportation departments face the possibility of being lost in the increasing volume of traffic details they handle. For example: assume the crash is detected on the highway that is being tracked using sensors installed on roadside poles by the transport authority. The directions should be given to traffic controllers to open an emergency exit road after investigating. This will require traffic to travel across the crash to prevent any road blockage. But the traffic controllers do not have advanced knowledge of any successful incident expected to take place at the emergency exit soon [3].

Text mining, also known as text analysis, is the process of transforming unstructured text into meaningful and actionable information. By identifying topics, patterns, and relevant keywords, text mining allows us to obtain valuable insights without needing to go through all data manually. Machine learning is an AI-derived discipline that focuses on developing algorithms that allow computers to learn from examples-based tasks. Machine learning models need to be equipped with input data, after which they can automatically predict with some degree of precision. The automatic text processing is proper when data mining and machine learning are merged.

The first thing should train a classifier subject model, by importing and marking a series of examples manually. The model should learn to distinguish topics and start creating correlations as well as its own predictions after being fed on many examples. To achieve reasonable standards of precision, you can feed a wide number of examples of the models that are representative of the problem you are trying to solve [6, 7].

Heterogeneity is the fundamental concern with accident data investigation is to recognize the most persuasive feature influencing accident recurrence and seriousness of the accident. The real issue with accident dataset analysis is its heterogeneous behavior. Heterogeneity in accident data is exceedingly undesirable and unavoidable [7]. With the growing amount of traffic information obtained from floating car data, it is highly beneficial to identify useful traffic patterns from the cumulative vast historical data collection, such as congestion patterns. Nevertheless, owing to the immense scale of the data collection, and the complexities and dynamics of traffic phenomena, it is difficult [8].

The accuracy of using traditional statistical analysis methods greatly relies on the size of the data. However, in many situations, data can be limited. The issue of small sampling has been always a problem in using crash data [9].

The proposed model integrated Variant algorithms in the preprocessing and text mining phases to enhance accuracy. It is not affected by the data size or quality. It uses efficient methods in preprocessing that enables the mining phase to work on any sample as presented the model applies on more one datasets (Maryland U.S. and Leeds UK).

The rest of this paper is organized as follows. Section 2 surveys the related works. Section 3 introduces the research objective. Section 4 discusses methods and techniques concerning text mining of the proposed model. Section 5 gives a description of the used datasets. Section 6 extends the experimental studies and results. Finally, Section 7 concludes the paper and presents suggestions for future work.

II. Related Work

Theoflatos et al. (2019) compared numerous machine learning approaches for real-time crash prediction for estimation of crashes. (including K-Nearest Neighbor, Naïve Bayes, Decision Tree, Random Forest, SVM, and Shallow Neural Network) and the Deep Feedforward Neural Network (DFNN) and found that the DFNN had more robust results in terms of different output parameters relative to other models [10].

Poch and Mannering used a seven-year incident dataset from 63 intersections in Bellevue, Washington (all based on procedural changes), this work tests a pessimistic binomial approximation of the recurrence of accidents at the approaches to crossing points. The calculation comes about disclosing crucial intersections between variables related to traffic and geography, and crash rates. The purpose of work-study offers exploratory analytical and objective data that may stimulate a way to work with the gage of the crash minimizing the benefits of multiple planned improvements on intersections that are operationally lacking [11].

Karlaftis and Tarko used the inquiry to aggregate the data and subsequently sorted the crash dataset into different categories and clustered output of the examined dataset by using Negative Binomial (NB) to classify the cause of the accident by a driver's centering age, which could have a few results [12].

Kwon OH used Naive Bayes and Decision Tree classification methodology to examine road safety-related factor dependencies [13]. Youthful Sohn utilized an alternative algorithm to enhance the accuracy of different types of classifiers for two severity categories of a traffic accident and every classifier utilized the neural network and decision tree [14].

Junhua Wanga used Two modeling approaches are implemented, including a conditional logistic regression model and a support vector machine model, and contrasted with forecasting crashes. The technique is evaluated based on the data obtained from the Shanghai Middle Ring Expressway, one of the major rings in Shanghai city, China [15].

Logistic regression was applied to collect crash-related information from traffic police reports, keeping in mind the end aim of examining the role of various situations in the incident severity. The demanded probity model was used to measure the effect on accident severity of pedestrian collision in a rural area of the highway and zone sort factors [16].

The model with the best fit and most elevated prescient capacity was used to classify the severity-related circumstances of the highway, an ecological problem, vehicle, and driver. Gadget usage, travel speed, impact intent, use of drugs and liquor, person condition, regardless of whether the driver is to blame, age, curve/grade, and rural/urban nature presence in the crash area were established as crucial elements for having an adverse effect on older drivers muddled in single-vehicle accidents [17].

Machine learning techniques such as the Support Vector Machine Algorithm (SVM) and the Fuzzy Clustering Models provide a modern and theoretically more reliable way to investigate the conventional safety issue. For example, in modeling the crash frequency on freeways, the negative binomial regression and the artificial neural network were compared and found the artificial neural network in modeling the crash frequency on freeways and found that the artificial neural network slightly outperformed the negative binomial regression [9].

In forecasting accident injury frequency on a mountainous highway based on real-time traffic and weather data, the analysis compared the fixed-parameter logistic model, the SVM model, and the random parameter logit model Other research suggested a new approach based on clustering algorithm and SVM model get a 78.0% accuracy [18].

Via Virtual Crash training tools, Oana Victoria Oțăt set out to educate the technological skills and analytical reasoning of the students in assessing the impacts of the pedestrian-vehicle. Therefore, familiarizing students with the methods offered by the program and to assist them in the study of virtual simulation of traffic accidents. Using the Simulated Crash program, students can learn how to assess the vehicle trajectory, based on vehicle movements during the pre-and post-impact phase [19].

III. Research Objective

The main objective of this work-study is to establish a model can achieve the optimum accuracy and identify the factors behind crashes or accident that could be helpful to reduce accident ratio in near future and could be helpful to save many lives, support in constructing the roads infrastructure, deteriorate wealth destruction as well as many other things. These outcomes will impact the urban movement police authorization measures, which will change the improper conduct of drivers and secure the minimum experienced street users. And give a fast response in case of crashes to get immediate support from EMS (Emergency Medical Service) unit for treatment. All the stakeholders and emergency teams will be automatically informed if an emergency is detected.

Proposes an IoT-based traffic management system for smart cities where traffic flows can be remotely managed by on-site traffic officers via their smartphones or centrally tracked or managed by the Internet [20]. There are circumstances that require defining policy objectives to constantly supply traffic information to the drivers, to keep the traffic moving.

IV. Proposed Model

The proposed Traffic Management Model discusses the main five characteristics as shown in Fig. 1 which implemented through this work-study using the Text Mining and Machine learning techniques.

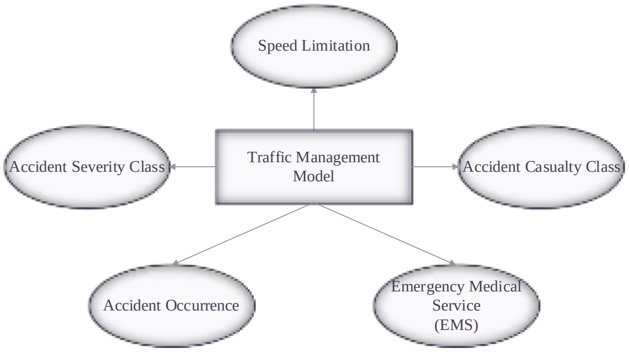


Fig. 1. Traffic Management Model Attributes.

The amount of data created is very huge and the amount of traffic information is collected through floating car data, so semantic models for data fusion and efficient algorithms for artificial intelligence are required to organize and process these data for extracting meaningful information especially from the accumulated massive historical dataset. Hence, data analyzing, and processing consider the big challenge for the data generated by IoT applications [2, 8].

Machine learning is a set of algorithms and statistical models that are used by computers to perform a required task. Machine learning can be used in traffic prediction. The data collected could be used in the construction of an idea display current traffic in the city and could be used in the future in making predictions of traffic & a congestion analysis can be done [21]. The explorer of knowledge sources that contain text or unstructured data is called “text mining”.

Text mining (also referred to as text analytics) is an artificial intelligence (AI) technology that transforms free (unstructured) text in documents and databases using natural language processing (NLP) into standardized, organized data suitable for analysis or to drive algorithms for machine learning (ML) [6].

For text mining, there are distinct approaches and techniques. The method of extracting essential patterns to discover information from textual data sources is text mining. Text mining is a multidisciplinary discipline focused on data retrieval, data mining, machine learning, digital linguistics, and statistics. It is possible to apply several text mining techniques to retrieve information, such as grouping, description, clustering, etc. Text mining deals with text in natural language and is encoded in semi-structured and unstructured format [22].

The structured data created by text mining can be integrated into databases, data warehouses, or business intelligence dashboards and used for descriptive, prescriptive, or predictive analytics. And the workflow detailed as following:

A. Data Source and Gathering

To make an effective model generate an accurate result, a real dataset from a trusted source must be provided, so the used datasets generated from official governmental source belongs to the US and UK government and published on the governmental official websites. A data preparation method, involving crash data filtering, floating car data filtering, and data matching on the road network, is introduced for the safety analysis purpose. And it contains:

- Dataset selection: The performance of real-time crash prediction relies greatly upon, in addition to the performance of the prediction model applied, a time-efficient and reliable data collection method [15].
- Data filtration and Cleansing: to define the most effective inputs (independent variables) compared to the traffic attributes in the related studies such as Weather condition, lighting, road conditions, … etc. and remove unwanted data which effect negatively on the accuracy and performance.

B. Data Preprocessing

Data preprocessing is performed on the road accident data to give it the proper shape required for analysis. Several attributes are transformed into a suitable form using data transformation methods such as Natural Language Toolkit (NLTK) and genism libraries.

Text clean-up: Removing any unnecessary or unwanted information such as ads from pages as the dataset (Text or CSV files) has a lot of noise and to perform machine learning algorithms efficiently as shown in Fig. 2.

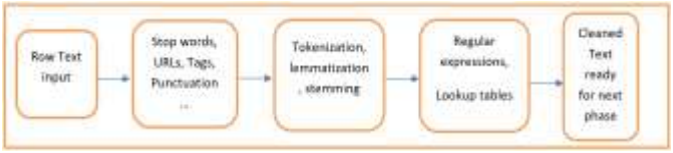

Row Text input
Stop words, URLs, Tags, Punctuation …
Tokenization, lemmatization , stemming
Regular expressions, Lookup tables
Cleaned Text ready for next phase

Fig. 2. Text Clean-Up.

1) Feature extraction (also called the collection of attributes): This is the method of characterizing the text to achieve a quantitative measurement package. The number of words in a text, word forms, syntactic details, for example. It is possible to use these features for further processing.

2) Categorical values handling: Text Mining can turn text into numbers in the most general words, many machine learning or data science operations may include text or categorical values in the dataset (basically non-numerical values). With numerical inputs, most algorithms perform better. Therefore, translating text/categorical data into numerical data and still making an algorithm/model to make sense of it is the biggest obstacle faced by analyst principal methods: One-Hot-Encoding and Label-Encoder. And are used to convert text or categorical data into numerical data which the model expects and performs better with.

3) Data transformation (feature scaling): (Data Normalization and standardization) since both the features have different scales, there is an opportunity that higher weightage is given to features with higher magnitude. this can impact the performance of the machine learning algorithm and clearly, no have to algorithm to be biassed towards one feature. The two most discussed scaling methods are Normalization and Standardization. Normalization typically means rescales the values into a spread of [0,1] Xchanged=(X−Xmin)/(Xmax−Xmin). Standardization rescales data to own a mean (μ) of 0 and variance (σ) of 1 (unit variance) Xchanged=(X−μ)/σ. Using sklearn.preprocessing and StandardScaler libraries.

C. *Index*

Creating an index of certain terms, their locations, and numbers. This allows quick access to and structuring of the processed data (Structured Database).

- Define the input features (independent variables) = X, and output labels (dependent variables) = y.
- Splitting the Dataset to (Training_Dataset) and (Testing_Dataset) to be X_train, X_test, y_train, y_test using sklearn.model_selection and train_test_split Libraries then define test_size data Percentage.

D. *Mining*

At this step, the text has been properly pre-processed and can now be ‘mined’. For that, applying different data exploration techniques to reveal new knowledge with high accuracy and performance.

- Dimensionality Reduction technique: PCA (Principal Component Analysis) is an unsupervised machine learning technique that attempts to derive a set of low-dimensional set of features from a much larger set while still preserving as much variance as possible. PCA can be thought of as a clustering algorithm. Usually, the original data is normalized before performing the PCA. It can be used for feature selection and visualizing higher-dimensional data where the feature pp is large. It is an unsupervised learning technique that can be used to identify patterns, clusters, and perhaps any latent information. Using sklearn.decomposition and PCA libraries. There are many techniques for the selection (reduction) such as wrappers but in choosing the most desirable subset of characteristics, PCA was concluded to be accurate by obtaining more variables of highest variance and low collinearity. Thus, in analysis and prediction, the right choice of algorithms offers performance. And it is more commonly used to reduce the dimensionality of a large dataset such that implementing machine learning where the original data is essentially high dimensional data becomes more realistic [23].
- Classification Techniques: Classification could be a form of supervised learning. It specifies the category to which data elements belong and is best used when the output has finite and discrete values. It predicts a category for an input variable yet. Classification predictive modeling involves assigning a category label to input examples.

E. *Result Analysis*

The mining steps produce raw results. These need to be evaluated and visualized so that they can be interpreted with respect to the questions the text-miner wants to investigate.

- Measurement of Accuracy: The classification accuracy is one of the important measures of how correctly a classifier classifies a record to its class value. The confusion matrix is an important dataa structure that helps in calculating different performance measures such as precision, accuracy, recall, and sensitivity of classification technique on some data using sklearn.metrics and confusion_matrix libraries. As shown in Table I.

TABLE I. CONFUSION MATRIX

	Negative	Positive
Negative	TN (True Negative)	FN (False Negative)
Positive	FP (False Positive)	TP (True Positive)

$$\text{Accuracy} = (TP + TN)/ (TP + TN + FP + FN) \quad (1)$$

$$\text{False Positive Rate} = FP/(TN+FP) \quad (2)$$

$$\text{Precision} = TP/(FP+TP) \quad (3)$$

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

$$\text{Recall} = TP/(TP+FN) \quad (5)$$

Using sklearn.metrics and accuracy_score, precision_ score, recall_score libraries in python to define the detailed results according to the previous functions.

F. *Emergency Medical Service*

To improve the possibilities of survival for passengers involved in car accidents, it is desirable to scale back the latent period of rescue teams and to optimize the medical and rescue resources needed. A faster and more efficient rescue will increase the probabilities of survival and recovery for injured victims. Thus, once the accident has occurred, it is crucial to managing the emergency rescue and resources efficiently and quickly. An Automatic Crash Notification system will automatically notify the closest emergency unit when a vehicle crash. These units will determine the character of the crash and, making it possible to predict the severity of injuries, data from vehicular sensors will allow the unit to judge if the vehicle has been involved during a collision. By using vehicular communications, cars involved in an accident can send an alert and other important information about the accident to nearby vehicles and to the closest wireless base station. Soon, a community-based effort involving the state departments, public organizations, and the industry is required to deploy the specified technology and infrastructure to attach all the vehicles on the road and therefore the emergency services. Basically, the data to be sent after an accident should include the following: (a) the time when the accident has occurred, (b) the placement of the vehicle to see the placement of the injured consistent with (longitude and latitude), (c) the characteristics of the vehicle and collision severity (allowing rescue services to send appropriate equipment to the accident site, and to warn them about the amount of complexity and dangers). All this information helpful in determining the severity of the impact, making it possible to avoid wasting lives, manage resources efficiently, and enable crashed vehicles to be far from the positioning, restoring traffic flow quickly.

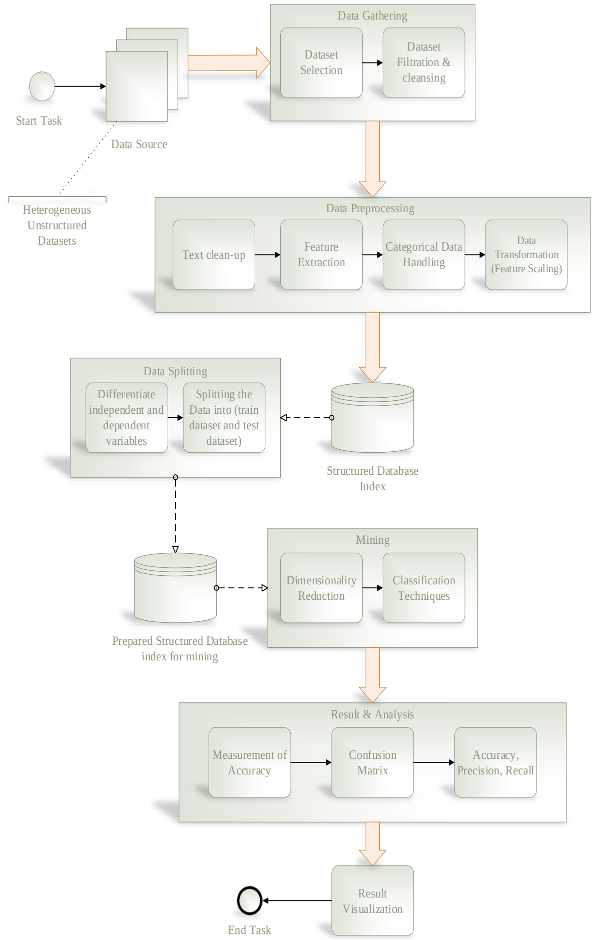

Start Task
Data Source
Heterogeneous Unstructured Datasets
Data Gathering
Dataset Selection
Dataset Filtration & cleansing
Data Preprocessing
Text clean-up
Feature Extraction
Categorical Data Handling
Data Transformation (Feature Scaling)
Data Splitting
Differentiate independent and dependent variables
Splitting the Data into (train dataset and test dataset)
Structured Database Index
Prepared Structured Database index for mining
Mining
Dimensionality Reduction
Classification Techniques
Result & Analysis
Measurement of Accuracy
Confusion Matrix
Accuracy, Precision, Recall
Result Visualization
End Task

Fig. 3. Traffic Management Model Workflow.

Accident notification systems will be specially designed for post-collision rescue services. New Intelligent Transportation Systems will emerge with the capability of improving the responsiveness of roadside emergency services. So, database to store the needed data must be designed, in used database the counties splitted to determine all the rescue units in the county related to the longitude and latitude, the rescue way according to the point either ground or air ambulance, the notified unit according to the nearest point in the crash location, and the treatment point which responsible to treat the injured persons according to injured severity. Using MySql with Python to implement that output. The detailed Traffic Management Model Workflow is shown in Fig. 3.

V. Description of Datasets

The traffic accident data is obtained from the online data source for Leeds UK. This data set comprises a massive number of accidents that occurred during a specific duration. Initial preprocessing of the data results in a set of attributes that found to affect the crash and the level of severity The attributes selected for analysis are several vehicles, time of the accident, road surface, weather conditions, lighting conditions, casualty class, the gender of casualty, age, type of vehicle, day, and month of the accident [24].

Crash data for Maryland U.S. from 2016 through 2019. Data is accurate as of the creation of the data. Only Approved Crash reports have been included in the file. Related datasets include Statewide Vehicle Crashes. This dataset provides general information about each collision and details of all traffic collisions occurring on the county and local roadways within Montgomery County, as collected via the Automated Crash Reporting System (ACRS) of the Maryland State Police, and reported by the Montgomery County Police, Gaithersburg Police, Rockville Police, or the Maryland-National Capital Park Police. All this data is unstructured in text format located in CSV files, the data contains a set of highly important attributes which are very powerful in Traffic Management Model: Quarter of the year, Light Condition, Road Junction, Collision Type, Road Surface Condition, Road structure Condition, Weather Condition, Vehicle Characteristics, Crash severity, Human Class and Characteristics, Speed Limit on the related road, Detailed Date and Time, Geographical Location and other attributes which may enhance the accuracy in this model.

The real-time activity patterns are linked with the historical data to predict how the patterns may evolve soon, which is indeed valuable for transportation management [5]. The crash severity model is concerned with predicting the distribution of crashes or injuries by severity, given that a crash has already occurred: the model does not predict crash probability itself.

VI. Experimental Study and Results Discussions

In this section, the detailed results presented regards the experimental studies and make a comparison between the different classifiers and each accuracy.

A. *Leeds UK Dataset (Casualty Class and Casualty Severity)*

As stated in Prayag Tiwari's studies the researcher applied three classifiers algorithms (decision tree, Naïve Bayes, and SVM), and according to that practical studies they achieved better accuracy by using clustering techniques such as (Self Organizing Map (SOM) and k-modes) based on casualty class to determine clearly that what circumstances affect and who is involved more in an accident between the driver, passenger, or pedestrian, and the better classifier in Prayag Tiwari's work-study is Decision Tree which achieved accuracy 81% better than the (Naïve Bayes, and SVM) and lowest classifier accuracy is SVM which achieved accuracy 75.58% as shown in Fig. 4 [25].

In this practical study, more classifiers applied to classify this dataset based on casualty class and severity class, these classifiers classified data into 3 classes for both. The output of this classifier is determined based on the precision, recall, error rate, and other various factors that play an important role in accuracy measurement. And based on casualty class and severity class, so the Specialists can see clearly that what circumstances affect and who is involved more in an accident between the driver, passenger, or pedestrian. In addition to Prayag Tiwari's Research Work elements, the severity class which is considered an added contribution selected for high traffic management perspective is added. The other contribution which considers an added value in this model is utilizing more classifiers algorithms which are (Random Forest and Logistic Regression). Already The better classifier in this work-study regards Casualty Class is Random Forest classifier which achieved accuracy 87.79% better than others and lowest classifier accuracy is Logistic Regression which achieved accuracy 75% and the Decision Tree is better classifier in this work-study regards Casualty Severity which achieved accuracy 88.02%, and lowest classifier accuracy is Naïve Bayes which achieved accuracy 86.35% as shown in Table II and Fig. 4 and 5.

TABLE II. TRAFFIC MANAGEMENT MODEL LEEDS UK DATASET RESULTS

Leeds UK (Casualty Severity)						Leeds UK (Casualty Class)				
Algorithm	**Accuracy**	**Precision**	**Recall**	**TP Rate**	**FP Rate**	**Accuracy**	**Precision**	**Recall**	**TP Rate**	**FP Rate**
Random Forest	87.12%	0.823	0.871	0.871	0.786	87.79%	0.88	0.875	0.875	0.126
Decision Tree	88.02%	0.88	0.88	0.88	0.22	87.75%	0.883	0.877	0.877	0.127
Naïve Bayes	86.35%	0.816	0.863	0.863	0.776	79.45%	0.792	0.795	0.795	0.189
SVM	87.81%	0.878	0.878	0.878	0.122	77.38%	0.7738	0.7738	0.774	0.226
Logistic Regression	86.85%	0.86846	0.86846	0.868	0.132	75%	0.75	0.75	0.75	0.25

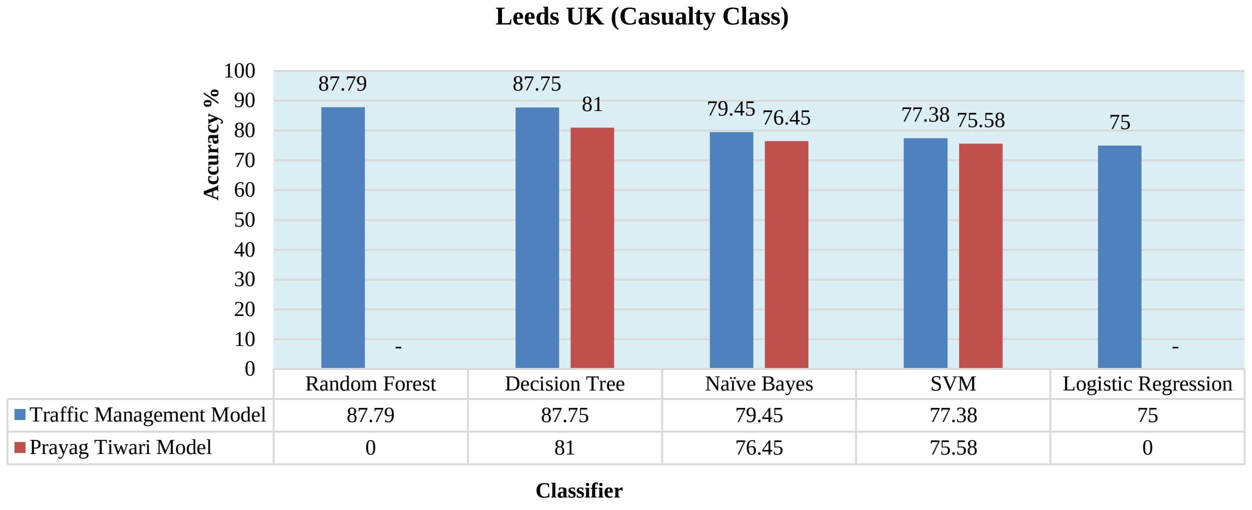


Fig. 4. The Accuracy Chart results between Traffic Management Model and Prayag Tiwari's Research Work.

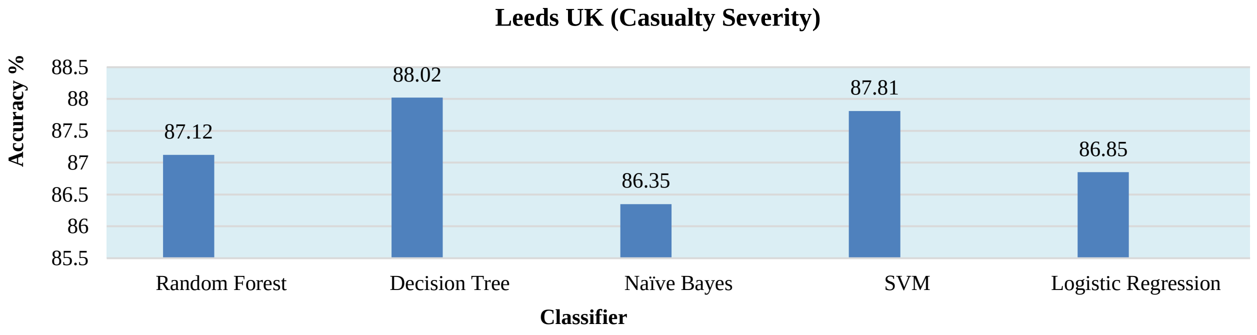


Fig. 5. The Accuracy Chart results in Leeds UK (Casualty Severity Classification).

B. Maryland U.S. Dataset (Speed Limitation)

Speed limit information is found to be very valuable in predicting crash occurrence. However, speed limits may have biased coefficients, most likely attributable to unobserved safety-related effects. So, it is highly significant to define the attributes which affect the speed limitation such as (Light Condition, Road Junction, Road Surface Condition, Road structure Condition, Weather Condition, Vehicle Characteristics, Geographical Location, and other attributes which may enhance the accuracy in this model). In these studies, using Maryland U.S. Dataset 5 classifier algorithms applied to define the most powerful algorithm in Speed Limitation and the Decision Tree Regression is the best algorithm to define the speed limit which achieved 98.93% accuracy and the lowest accuracy achieved by Logistic Regression with accuracy 63.91% as shown in Table III and Fig. 6.

TABLE III. MARYLAND U.S. DATASET (SPEED LIMITATION)

Maryland U.S. (Speed Limitation)					
Algorithm	Accuracy	Precision	Recall	TP Rate	FP Rate
Random Forest	88.24%	0.88235	0.882	0.882	0.118
Decision Tree Classification	90.59%	0.90588	0.905	0.906	0.094
SVM	82.35%	0.82353	0.823	0.824	0.176
Decision Tree Regression	98.93%	0.9893	0.989	0.989	0.011
Logistic Regression	63.91%	0.63905	0.639	0.639	0.361

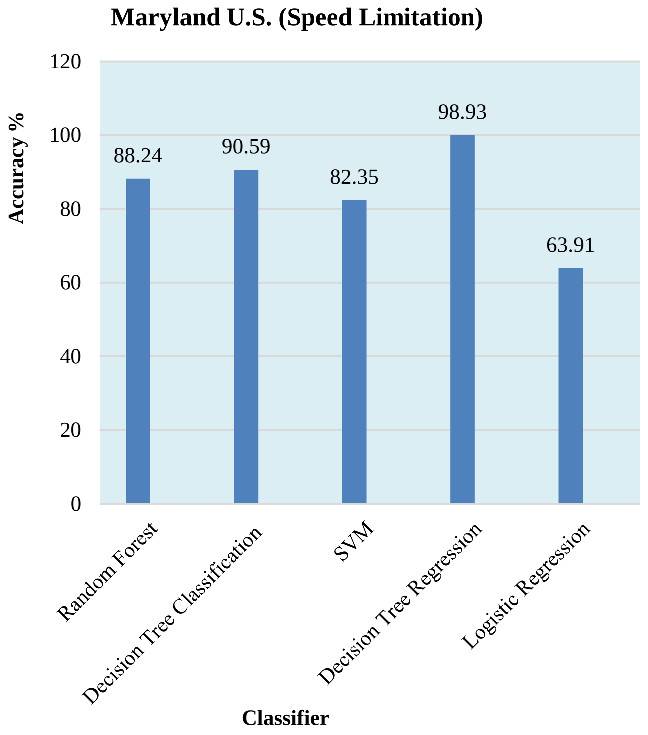


Fig. 6. The Accuracy Chart results in Maryland U.S. (Speed Limitation).

C. Maryland U.S. Dataset (Casualty Class)

According to these practical studies in Maryland U.S. Datasets using more attributes than defined in Leeds UK Datasets to achieve more accuracy and already better accuracy achieved based on Casualty Class so, the Specialists can see clearly that what circumstances affect and who is involved more in an accident between the driver, passenger or pedestrian, and the better classifier in this work-study is Decision Tree Classification which achieved accuracy 89.38% better than the others classifiers used, and lowest classifier accuracy is Logistic Regression which achieved accuracy 83.08% as shown in Table IV and Fig. 7.

TABLE IV. MARYLAND U.S. DATASET RESULT (CASUALTY CLASS)

Maryland U.S. (Casualty Class)					
Algorithm	**Accuracy**	**Precision**	**Recall**	**TP Rate**	**FP Rate**
Random Forest	88.54%	0.8854	0.8854	0.885	0.115
Decision Tree Classification	89.38%	0.89377	0.89377	0.894	0.106
SVM	86.65%	0.86646	0.86646	0.866	0.134
K-Nearest Neighbor	84.90%	0.849	0.849	0.849	0.151
Logistic Regression	83.08%	0.8308	0.8308	0.831	0.169

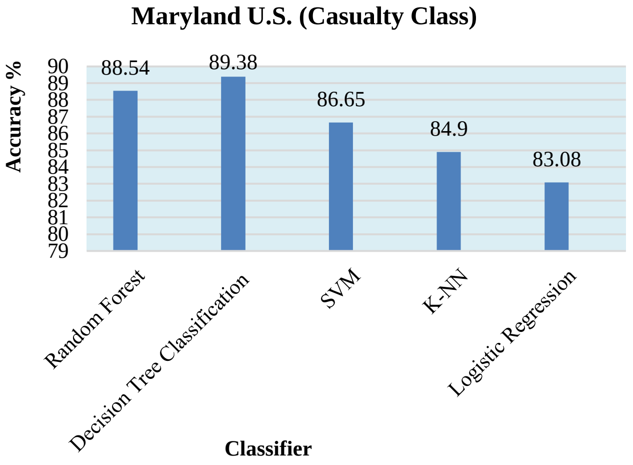


Fig. 7. The Accuracy Chart results in Maryland U.S. (Casualty Class).

D. Maryland U.S. Dataset (Destruction Severity Class)

The accident severity classification is a significant factor in traffic management especially in Emergency Medical Service to define the type of rescue tools and the fast response for support to the Casualty victims. On the other side the accident severity classification has an important impact on road traffic congestion and the more serious the accident, the greater the traffic congestion and it supports the traffic management authorizers for urgent decision making such as preparing alternative routes or preparing the emergency roads. In these studies, using Maryland U.S. Dataset five classifier algorithms applied to define the most powerful algorithm in

Crash Severity Classification and the Random Forest is the best algorithm to define the Destruction severity Class which achieved 98.63% accuracy and the lowest accuracy achieved by Logistic Regression with accuracy 80.8 % as shown in Table V and Fig. 8.

E. Maryland U.S. Dataset (Accident Occurrence)

Predicting the occurrence of crashes helps study safety traffic planning and make improvements in the traffic elements. Real-time crash prediction helps identify and prevent crashes before they happen. Therefore, it has become a hot topic in the ITS industry, both crash and non-crash events are needed in crash prediction [15]. So, the proposed traffic management model applied in Maryland U.S. datasets to define the factors which contributed to the accident in this study four classifier algorithms applied to define the most powerful algorithm in Accident Occurrence prediction and the better classifier in this work-study is SVM which achieved accuracy 98.84% better than the other classifier used, and lowest classifier accuracy is K-Nearest Neighbor which achieved accuracy 98.26% as shown in Table VI and Fig. 9.

TABLE V. MARYLAND U.S. DATASET RESULT (DESTRUCTION SEVERITY CLASS)

Maryland U.S. (Destruction Severity Class)					
Algorithm	**Accuracy**	**Precision**	**Recall**	**TP Rate**	**FP Rate**
Random Forest	98.63%	0.98634	0.98634	0.986	0.014
Decision Tree Classification	98.48%	0.9848	0.9848	0.985	0.015
SVM	89.61%	0.89605	0.89605	0.896	0.104
K-Nearest Neighbor	94.76%	0.94764	0.94764	0.948	0.052
Logistic Regression	80.80%	0.80804	0.80804	0.808	0.192

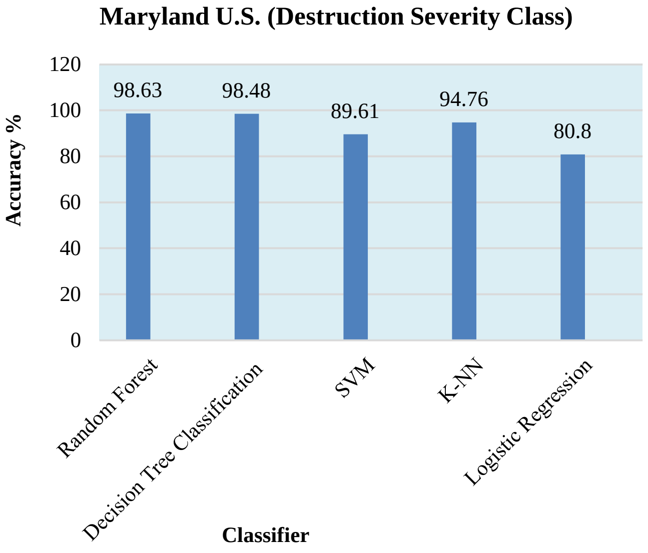


Fig. 8. The Accuracy Chart results in Maryland U.S. (Destruction Class).

TABLE VI. MARYLAND U.S, DATASET RESULT (ACCIDENT OCCURRENCE)

Maryland U.S. (Accident Occurrence)					
Algorithm	**Accuracy**	**Precision**	**Recall**	**TP Rate**	**FP Rate**
Random Forest	98.53%	0.9853	0.9853	0.985	0.015
Decision Tree Classification	98.71%	0.9871	0.9871	0.987	0.013
SVM	98.84%	0.9884	0.9884	0.988	0.012
K-Nearest Neighbor	98.26%	0.9826	0.9826	0.983	0.017

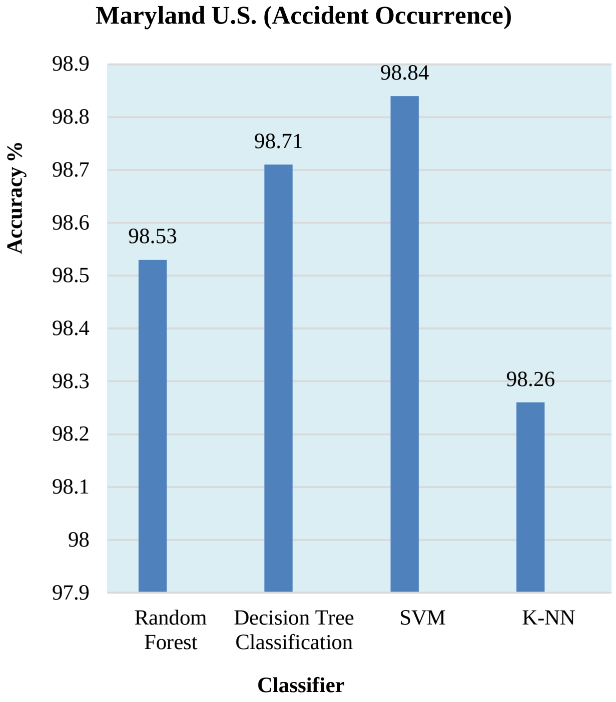


Fig. 9. The Accuracy Chart results in Maryland U.S. (Accident Occurrence).

VII. CONCLUSION AND FUTURE WORK

In these practical studies, there are several Text Mining approaches has been performed to analyze different real datasets which supported by governmental agencies to be trusted data source such as Leeds UK, using different classification and machine learning techniques. One of the contributions is this work-study outperforms Prayag Tiwari's Work-study and better results achieved using more classifiers with higher accuracy from this way based on casualty class and casualty severity so the Specialists can see clearly that what circumstances affect and who is involved more in an accident between the driver, passenger, or pedestrian. And to which degree the victim was affected by the accident. On the other side using Maryland U.S. datasets the previous model applied to classify these datasets into traffic management characteristics needed to define the aspects which have an important role in traffic management processes such as (Accident Casualty Class, Accident Severity Class, Speed Limitation across Geographical Locations, Accident Destruction Class, and Accident Occurrence prediction). Multi algorithms applied to get the optimum classifiers used in this model such as (Random Forest, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Logistic Regression, and Naïve Bayes). To improve the chances of survival for victims involved in road accidents, it is significant to reduce the response time of rescue teams and to optimize the medical and rescue resources needed. So, the proposed Emergency Medical Service (EMS) once the accident has occurred, it is crucial to managing emergency rescue and resources efficiently and quickly. In future work, the researcher recommends establishing a framework that saves all data to study the usual behavior for each user to further analysis to be used in IoT road traffic applications.

ACKNOWLEDGMENT

I sincerely thank all those who helped me in completing this task, especially the participant supervisors.

REFERENCES

[1] 1Suraj Kumar G Shukla, 2Aadithya Kandeth, 3D. Sai Santhiya, 4Kayalvizhi Jayavel SRM Institute of Science & Technology, Kattankulathur, Chennai "Efficient Traffic Management System" International Journal of Engineering &Technology, 7 (3.12) (2018) 926 -932.

[2] Luci Sumi, and Virender Ranga. National Institute of Technology, Kurukshetra "Sensor enabled Internet of Things for Smart Cities" 2016 Fourth International Conference on parallel, Distriputed, and Grid Computing (PDGC). Conference Paper · December 2016. DOI: 10.1109/PDGC.2016.7913163.

[3] Deepti Goel, Santanu Chaudhury, and Hiranmay Ghosh. 2017. "An IoT Approach for Context-aware Smart Traffic Management Using Ontology". In Proceedings of WI '17, Leipzig, Germany, August 23-26, 2017, 8 pages. DOI: 10.1145/3106426.3106499.

[4] Dovydas Skrodenis, Vilnius Gediminas Technical University, Road Research Institute Saulėtekio al. 11, 10223 Vilnius, Lithuania "Road Traffic Management During Special Events" Conference paper, First Online: 18 June 2019 pp 104-109.

[5] Sasan Amini, Eftychios Papapanagiotou, and Fritz Busch, Chair of Traffic Engineering and control, Technical University of Munich "Traffic Management for Major Events" Digital Mobility Platforms and Ecosystems. DOI: 10.14459/2016md1324021.

[6] Sonali Vijay Gaikwad, Archana Chaugule, and Pramod Patil, "Text Mining Methods and Techniques". International Journal of Computer Applications (0975 – 8887), Volume 85 – No 17, January 2014.

[7] Karlaftis M, Tarko A (1998) "Heterogeneity considerations in accident modeling". Accid Anal Prev 30(4):425–433.

[8] Lin Xu*, Yang Yue, Qingquan Li, State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, "Identifying Urban Traffic Congestion Pattern from Historical Floating Car Data". 13th COTA International Conference of Transportation Professionals (CICTP 2013), Procedia - Social and Behavioral Sciences 96 (2013) 2084 – 2095.

[9] Guo, Y., Graber, A., McBurney, R.N., Balasubramanian, R., 2010. "Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms". BMC Bioinformatics 11 (447), 19p.

[10] Theoflatos, A., Chen, C., Antoniou, C., 2019. "Comparing machine learning and deep learning methods for real-time crash prediction". Transportation Res. Record 1–10.

[11] Poch, Mark, and Fred Mannering. "Negative binomial analysis of intersection-accident frequencies." Journal of transportation engineering 122.2 (1996): 105-113.

[12] Karlaftis M, Tarko A (1998) "Heterogeneity considerations in accident modeling". Accid Anal Prev 30(4):425–433.

[13] Oh HoonKwon, WonjongRhee˙, YoonjinYoon, "Application of classification algorithms for analysis of road safety risk factor dependencies". Accident Analysis & Prevention, Publisher: Elsevier, Date: February 2015. DOI: 10.1016/j.aap.2014.11.005.

[14] So Young Sohn , Sung Ho Lee "Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea". February 2003 Safety Science 41(1):1-14, DOI: 10.1016/S0925-7535(01)00032-7.

[15] Junhua Wanga, Tianyang Luoa, Ting Fua,b,∗ "Crash prediction based on traffic platoon characteristics using floating car trajectory data and the machine learning approach", ON, N2L 3G1, Canada, ELSEVIER.

[16] Ali S. Al-Ghamdi, "Using logistic regression to estimate the influence of accident factors on accident severity" [J]. Accident Analysis and Prevention, 2002: 729-741.

[17] Sunanda Dissanayake, Jian John Lu. "Factors influential in making an injury severity difference to older drivers involved in fixed object-passenger car crashes" [J]. Accident Analysis and Prevention, 2002: 609-618.

[18] Yu, R., Abdel-Aty, M., 2014. "Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data". Saf. Sci. 63, 50–56.

[19] Oana Victoria Oțăt University of Craiova, Faculty of Mechanical Engineering, 107 Calea Bucuresti Str., Craiova, Romania, otatoana@yahoo.com "traffic management training via dedcated software" The European Proceedings of Social & Behavioural Sciences EpSBS, ISSN: 2357-1330 , https://doi.org/10.15405/epsbs.2019.08.03.184.

[20] 1Syed Misbahuddin, 2Junaid Ahmed Zubairi, 3Abdulrahman Saggaf, 4Jihad Basuni, 5Sulaiman A-Wadany and 6Ahmed Al-Sofi, "IoT Based Dynamic Road Traffic Management for Smart Cities", Fredonia NY 14063 Conference Paper · December 2015.

[21] Sagarika Verma, Sayali Badade Computer Science and Engineering, MIT-ADT University, Pune, India. "Traffic Prediction Using Machine Learning" Proceedings of National Conference on Machine Learning, 26th March 2019. ISBN: 978-93-5351-521-8.

[22] Ramzan Talib∗, Muhammad Kashif Hanify, Shaeela Ayeshaz and, Fakeeha Fatimax "Text Mining: Techniques, Applications and Issues" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7 No. 11, 2016.

[23] Vinayak Hegdea, Shruthi G. Kinia, and Sahana A, "A Comparative Study of Principal Component Analysis vs Wrapper Method, an Overview of Dimensionality Reduction Techniques Applied in Developing an Undergraduate Student Dropout Model", International Journal of Control Theory and Applications, Volume 9 • Number 42 • 2016, ISSN: 0974–5572.

[24] Prayag Tiwari University of Padova" Accident Analysis by using Data Mining Techniques" Thesis ·June 2017 DOI: 10.13140/RG.2.2.20091.41766/.

[25] Prayag Tiwari University of Padova, Sachin Agnihotri South Ural State University, Denis Kalitin National University of Science and Technology MISIS, "Road-User Specific Analysis of Traffic Accident Using Data Mining Techniques", Conference Paper · September 2017, DOI: 10.1007/978-981-10-6430-2_31.

Detection of nCoV-19 from Hybrid Dataset of CXR Images using Deep Convolutional Neural Network

Muhammad Ahmed Zaki[1], Sanam Narejo[2]*, Sammer Zai[3], Urooba Zaki[4], Zarqa Altaf[5], Naseer u Din[6]
Department of Computer System Engineering, Mehran University of Engineering and Technology, Jamshoro, Pakistan[1,2,3,5,6]
Institute of Information and Communication Technology, University of Sindh, Jamshoro, Pakistan[4]

***Abstract*—The Corona-virus spreads too quickly among humans and reaches more than 72 million people around the world until now. To avoid spread, it is very important to recognize the individuals infected. The Deep Learning (DL) technique for the detection of patients with Corona-virus infection using Chest X-rays (CXR) images is proposed in this article. Besides, we show how to implement an advanced model for deep learning, using Chest X-rays (CXR) images, to identify COVID-19 (nCoV-19). The goal is to provide an intellectual image recognition model for over-stressed medical professionals with a second pair of eyes. In using the current publicly available COVID-19 data-sets we emphasize the challenges (including image data-set size and image quality) in developing a valuable deep learning model. We suggest a pre-trained model of a semi-automated image, create a robust image data-set for designing and evaluating a deep learning algorithm. This will provide the researchers and practitioners with a solid path to the future development of an improved model.**

Keywords—COVID-19; artificial intelligence; deep learning; chest x-ray image analysis; convolutional neural network; InceptionV3.

I. Introduction

COVID-19 is a population affected pandemic that has continued to have catastrophic consequences worldwide. Individual screening is one of the key steps that can be used to manage the spread of infection at infected hot spots and to monitor the health of serious patients. The need for fast testing highlights simple and effective screening methods that are available. The novel COVID-19 (nCoV-19) is been called when viewed under an electron microscope because of its distinctive solar corona (crown-like) appearance [1]. As of December 2019, the epidemic of the nCoV-19 in Wuhan, China has an extent briskly to other nations [2]. The transmission from animals to human beings [3] is shown in Fig. 1. Infectious diseases caused by these types of viruses were identified as nCoV-19 by the World Health Organization (WHO) on Feb 12, 2020 [4]. China had received about 90,000 confirmed cases till (March 21, 2020) and the worldwide confirmation of more than seventy-two million cases [5] shown in Fig. 2(a), and total COVID-19 deaths are shown in Fig. 2(b). All the approaches under existing AI techniques detect pneumonia from COVID-19 by the use of the database through X-ray images. Among the difficulties, AI faces when it comes to detecting pneumonia is "how the machine knows that COVID-19 triggers the identification of pneumonia in the chest x-rays". The researchers also accept that there are complications, though majorities of deaths from nCoV-19 of vulnerable patients are due to pneumonia, "Dr. Tom Naunton Morgan Chief Physician at behold.ai. [6] said. There are a variety of pathogens that may potentially endanger the existence of pneumonia, including direct or indirect COVID-19 infections". The algorithms are preparing for real-time detection of pneumonia. In the majority of cases, the main cause of pneumonia is bacteria. The signs are cold, grip, and affected lungs. Soil and bird drops are also a cause of potential pneumonia. Some of the viruses that cause resentment and influenza lead to pneumonia. The justification for explaining and triggering these forms of pneumonia is to explain the problems by using a robust Artificial Intelligence (AI) program. AI can assist the public in various areas, including early cautions, diagnosis predictions, therapies, monitoring classification, treatments, analysis tools, and social control. A Canadian AI platform, the BlueDot [7] has proved its value and has gained some great reputation. Conferring to the AI model, the system was learned of an epidemic on 31st December 2019 earlier the announcement from the WHO on 9 January 2020. In January Journal of Travel Medicine [8], several researchers have collaborated with BlueDot. They mentioned that Wuhan China's virus was spread through travelers in 20 different cities worldwide. While BlueDot is a powerful AI tool, it has also been exaggerated by several forums. HealthMap [9] was the second AI model, which also declared an early alarm in Boston Children's Hospital, U.S.A., a scientist predicted that a COVID-19 significant outbreak after half an hour. The AI model reacted more quickly than a human being. Exact identification can save many lives with short computational power. The spread and development of the disease can also be managed via a well-trained model. United Nations (UN) Global Pulse scientists have analyzed several COVID-19 applications based on AI's. It was claimed that both Chest X-rays (CXR) and Computed Tomography (CT) scans can be used for COVID-19 detection built on AI modeling. The scholars also proposed that CT images for the identification of COVID-19 should be scanned using a cell phone [10].

*Corresponding Author

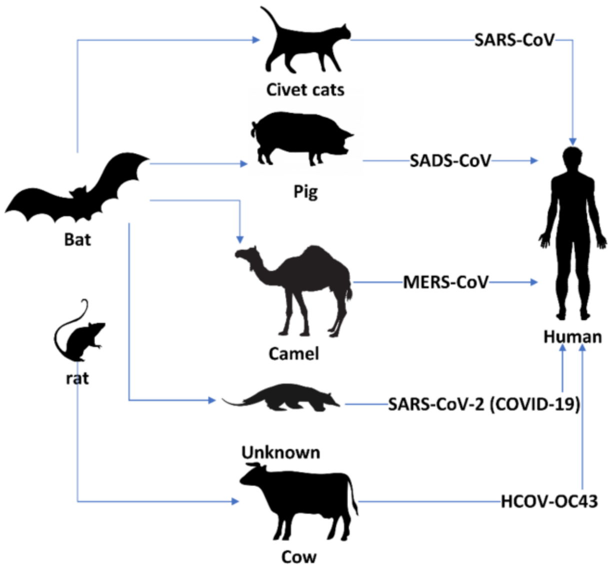


Fig. 1. Spread of nCoV-19 from Animals to the Human being Spread of nCoV-19 from Animals to the Human being.

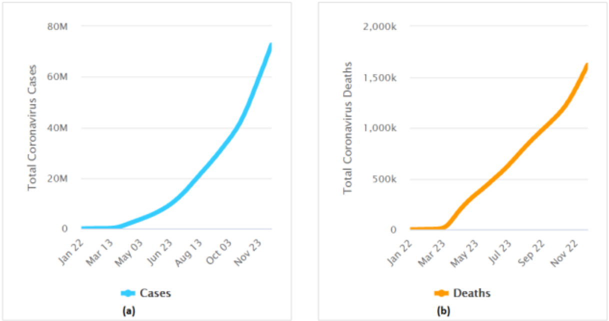


Fig. 2. Number of Cases and Deaths due to nCoV-19 Worldwide (a) Total Cases Worldwide (b) Total Deaths Worldwide.

AI will also provide ample help for the prediction and monitoring purposes to decide how long this COVID-19 pandemic is spread. Following the last pandemic in 2015, an AI-based system for the prediction of disease propagation is being developed for Zika viruses [11]. For COVID-19, these existing models can be used so that the system can be retrained with COVID-19 data. nCoV-19 data may also retrain the algorithm for the prediction of seasonal flu [12]. We need a great deal of training data for deep learning architecture for this we use different types of available datasets [13] [14]. The databases used are very small in state of the art methods. The enactment of a system that is trained with 60-120 images with just one particular type of pneumonia that has occurred due to nCoV-19 cannot be dependent. To improve the AI-based system performance for detecting nCoV-19, a huge dataset of CXR images is essential for multiple categories of pneumonia.

AI leads to the exploration of potential emerging drugs well. Taking into account the COVID-19 outbreak, various research laboratories indicated that AI had to look for the treatment vaccine. Scientists consider that the procedure of the AI model can speed up the COVID-19 search and vaccination process, so Google Deep Mind predicts the protein structures of COVID-19 and can provide useful vaccine discovery information. On the website of Deep Mind, it is also mentioned: We highlight that these predictions of the structure were not experimentally confirmed. We cannot be sure that the systems we provide are correct [15]". A refined, visualized evidence on the facts of COVID-19 is delivered by the data dashboard. The AI-based data for pursuing and predicting the nCoV-19 outbreak has been provided with various lists from MIT technologies, including HealthMap [16], NextTrain [17], and Upcode [18]. These dashboards offer a global opinion of the outbreak of COVID-19 in every country. The skimming of people in congested areas, or possibly affected areas, can be monitored by AI so that the body temperature of humans for the virus forecast against nCoV-19 was detected by an infrared camera in the Chinese railway station [19]. AI is used to manage the pandemic by scanning and implementing social separation and lockdown. As the South China Morning Post describes, Infrared cameras can scan the crowd for high temperatures in airports and railway stations throughout China. China uses a facial recognition system that can identify people with high body temperature and whether they are wearing a surgical mask or not. The positive COVID-19 patients can be significantly confirmed by imagery methods like the CT and CXR [20]. In recent studies to classify COVID-19, CT images for lung and soft tissue were examined. However, the downside of CT imaging is the expense scan and a high dose of the patient. Conversely, in every hospital and clinic, CXR is available to generate 2-dimensional (2D) thorax projection images [21]. The CXR model is generally the primary option to identify chest pathology and has been applied by radiologists in a limited number of patients to confirm COVID-19. The emphasis of this research is therefore only on using the chest x-rays imaging method for possible patients with nCoV-19 [22]. In CXR images, however, soft tissue with a bad contrast cannot readily be detected [23]. Computer-Aided Diagnoses (CAD) have been developed to help clinicians identify and measure distrusted infections of vital tissues automatically in CXR images to overcome these limitations. CAD systems are focused essentially on the fast progress of computing equipment such as graphical processing units (GPUs) to run algorithms for the processes of medical images, including improvement of images, segmentation of organ tumors, and interventional navigation [24]. Deep Neural Networks are one of the influential architectures of DL and has become used intuitively in multiple applications [25]- [27]. So far, the use of DL techniques in chest x-ray to classify and identify nCoV-19 is still very limited. Data is the first step in the creation of a diagnostic method in the sense of a COVID-19 pandemic. There are broad public collections of more CXR, also nCoV-19 CXR are collected. In this research, we isolate the public database of CXR pneumonia cases, in precise COVID-19 cases.

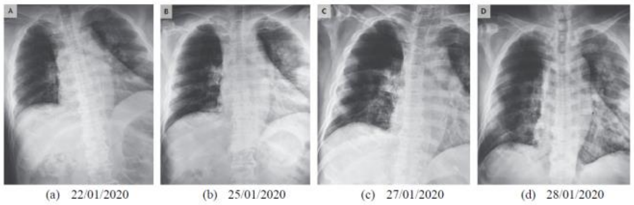


Fig. 3. 55-Year-Old Woman Survived from nCoV-19 [28].

Data must be obtained from public databases so as not to violate the confidentiality of patients. Fig. 3 shows an example of an infected female of age 55-year-old who survived COVID-19. This will provide important knowledge for the development and training of a deep learning system. These tools can be built to detect the characteristics of nCoV-19 concerning other pneumonia types. The purpose of this research is, therefore, to suggest a system for pre-trained DL classifiers, Convolutional Neural Network (CNN) as an advanced way of supporting x-ray images to analyze nCoV-19 automatically. The article is arranged according to the following. Section 2 describes the existing deep learning image classifiers. Section 3 delivers an overview of the related work. The suggested CNN model is defined in detail in Section 4. Section 5 provides trial results and discusses the performance of the model. The main prospects for this study are ended in Section 6.

II. Deep Learning Image Classifiers

One of the significant objectives of this study was to accomplish a state of the art grading outcomes using publicly handy data and models, with transfer learning to balances the limited sample size and speed up training processes so that modest hardware can provide reasonable results. In this section, we define some of the deep learning classifiers that are available today.

A. VGG19

(VGG) Visual Geometry Group was created based on the CNN architecture by Oxford Robotics Institute's Andrew Zisserman and Karen Simonyan [29]. It was presented at the Large Scale Visual Recognition Challenge in 2014. On the ImageNet data-set. To improve its image extraction, VGGNet uses small filters of 3×3, as compared to AlexNets with 11×11 filters. This deep network architecture is made up of two versions: VGG19 and VGG16, each with diverse layers and depths. VGG19 is more profound than VGG16. However, the numeral of parameters is greater for VGG19, and thus costlier to train the network than for VGG16.

B. DenseNet121

There are several important benefits for the Dense Convolutional Network they reduce the vanishing-gradient problem, increase the propagation of features, encourage the reuse of features, and significantly lower the number of parameters [30]. DenseNet121 is a 121-layer Dense Network interface that loaded the ImageNet database with pre-trained weights.

C. InceptionV3

The network consists of 159 layers and secured the 2014 ImageNet challenge with a top 5 accuracy of 93.3% [31]. Late versions are stated as Inception VN. N is the version number so InceptionV1, V2, and V3. The InceptionV3 network Implementation has many building blocks regular and irregular, with separate divisions of convolutions, mean, max pooling, concatenated, dropouts, and fully connected layers.

D. ResNetV2

To achieve strong convergence patterns, He et al., [32] have established Residual Neural Network prototypes by using skip connections to hop over certain layers of the network. The improved ResNet version is known as ResNet-V2. Even if the ResNet looks like the VGGNet, but it is about 8 times deeper than VGGNet [33].

E. Inception-ResNet-V2

The network is consisting of 572 layers deep, which combines the architecture of Inception with residual connections. Inception-ResNet-V2 is an InceptionV3 [34] variant. Over a million images in the ImageNet dataset are trained in Inception-ResNet-V2.

F. Xception

Xception model design is a linear heap of depthwise divisible convolution layers with residual connections that allow the deep-network design to be easily described and modified [35]. The Xception is an enhanced design of the Inception framework that substitutes standard initial units with distinctive depth convolutions.

G. MobileNetV2

CNN architecture for restricted computing power devices such as smartphones [36]. Sandler et al., [37] proposed the MobilleNetV2 model. MobileNet achieves this primary advantage by reducing the number of learning constraints and intuitively reducing memory consumption by inverted residuals using the linear bottleneck blocks. Besides, the pre-trained execution of MobileNetV2 is extensively available in many standard deep learning environments.

III. Related Work

A clear diagnosis and the cause of illnesses are identified, a major obstacle for doctors to reduce patient distress remains in time. Certainly, the usage of Image Processing (IP) and DL methods in biomedical image processing and analysis has delivered very satisfactory results. A brief overview of a few significant contributions from the existing literature is provided in this section.

Sethy et al. [38] suggested the identification of nCoV-19 based on the Support Vector Machine (SVM) and deep features using X-ray images. They had collected CXR images from the repository of Kaggle, GitHub, and Open-I repository. They mined the deep features of CNN prototypes and fed each individually to the SVM classifier. They have got an accuracy of 98.66%. with ResNet50 plus SVM.

Shan et al. [39] aim to estimate COVID-19 in CT scan by using the DL model named VB-Net. They used 300 images for validation and 250 images for training. They achieved a precision of 91.6%.

Butt et al. [40] suggested a model for detection from influenza-A viral pneumonia nCoV-19 with the use of deep learning techniques in pulmonary CT images. The CNN model provided 86.7% accuracy for CT images.

Wang et al. [41] used CT images for nCoV-19. The Transfer- Learning Model was also used to construct the algorithm with 89.5% accuracy reported.

Bhandary et al. [42] proposed the framework for the diagnosis of pneumonia and cancer. Two separate DL techniques were proposed. The first CXR images were classified with the help of SVM into a normal and pneumonia class and their performance was validated with additional pre-trained, deep learning models (ResNet50, VGG19, VGG16, and AlexNet). The second introduced a combination of handmade and studied features in the MAN to boost the precise rating during the lung cancer test.

Stephen et al. [43] suggested a new classifying pneumonia detection based on the ConvNet model trained from scratches based on a data-set from a collection of CXR images. The results obtained were 12.88% of training loss, 95.31% of training accuracy, and 93.73% of validation.

Ayan et al. [44] implemented an early diagnostic system based on the Xception and VGG16 CNN model. The study used 5856 frontal CXR images of the Kermany data-set. Test results show that the VGG16 network is better than the Xception network by precision 86%, sensitivity 85%, and recall 94%. The VGG16 network is more effective for classifying CXR images than the Xception network.

Varshni et al. [45] Suggested pre-trained ConvNet models (ResNet50, DenseNet-121, Xception, VGG-16, and DenseNet-169) with feature extractors tailed by diverse classifiers (K-nearest neighbors (KNN), SVM, Naïve Bayes, and Random Forest) for classifying abnormal and normal CXR pneumonia images.

In this article, we presented the CNN model of DL for classifying nCoV-19 from CXR images. To input CXR images into the CNN, a classifier is then used to set the outputs of the consequences of the classification.

IV. Research Methodology

Use Methods of deep learning (DL), with state of the art Computer Vision (CV) and IP, have recently been shown to provide enormous potential [46]. These technologies have been employed in various methods for the segmentation, recognition, and classification of high-performance medical imaging [47]. Some DL techniques include identification of skin cancer, breast cancer identification, classification, detection of lung cancer, etc. While these methods have shown great achievement in medical imaging, they involve a huge amount of data which in this area of applications is not yet available.

To identify nCoV-19, we implement a simple CNN model consisting of a convolution layer including 5×5 filter tailed by batch normalization layer, rectified linear unit (ReLU), a completely connected layers, SoftMax layer, and an output layer. The loss function is used to initialize the weights and cross-entropy in the classification layer [48]. Fig. 4 gives details of the CNN model.

A. Input Layer

This layer is liable for reading a pre-processed image data-set. Since the medical image contains several letters and medical symbols. Image sizes are varied as images are taken from different sources. Therefore, changes the input image size by 255×255. We cropped the region of the lung and chest as much as possible so it does not contain any extra region.

B. Convolution Layer

This layer is the critical layer in our suggested CNN model which will perform most calculations. This layer's chief function is to recover features from the image data-set and to preserve the spatial association between pixels. The functions are obtained using a series of filters, where an individual filter is designed based on the size of a 5×5 filter.

C. Batch Normalization Layer

This layer represents an extremely deep technique for NN training that stabilizes the convoluted feature values. The aim of using this layer is to decrease the number of training epochs essential for deep network creation and to stabilize the learning cycle.

D. ReLU Layer

This layer aims to substitute the negative values of a pixel with zero in the convolved features. This creates the non-linearity plan of the CNN network features.

E. Fully Connected Layer

All activation functions of the preceding layer are related to the neurons of this layer. The main task of this layer is to categorize the collected features in the specified classes from the image data-sets.

F. SoftMax Layer

This layer is merely used to consider the possible values of the previous layer activation function. The values can be interpreted in two groups of '0' and '1' in the diagnosis case.

G. Output Layer

The last layer of the CNN model can be labeled with the outcomes of the preceding layer. For example, '1' is marked at CoV-19^{+}, and '0' is marked at none of CoV-19^{-}.

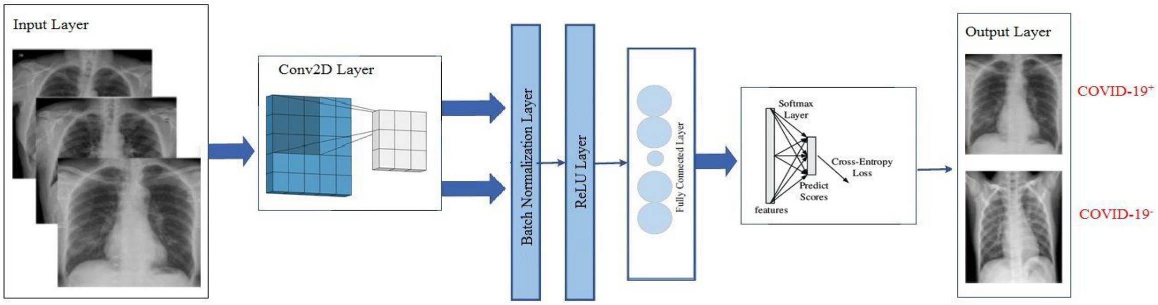


Fig. 4. CNN Model to Analyze nCoV-19 Disease.

V. Implementation Details

70% of Chest X-rays (CXR) images containing abnormal and normal cases are arbitrarily selected for CNN training to test the efficiency of the suggested DL classifier. Training parameters for deep convolution neural network (DCNN) architecture in this study includes initial learning rate=3 $\times 10^{-3}$ to accomplish the desired merging on this slight image data-set with few iterations and also to avoid the degradation issue as possible. The proposed CNN model has been trained with the Adam optimizer on the ImageNet data-set. For training purposes: the minimum learning rate = 4.78×10^{-9}, the number of epochs = 500, the batch size = 16, factor = 0.3, and patience = 1 all these hyper-parameters were used. The successively times of all DL models are fairly undersized ranging starting from 410 to 2845 seconds, due to the use of powerful GPU tools with a small chest X-ray (CXR) image data-set. On 10 images tested, the test times of the proposed model that resulted did not exceed 10 seconds. Finally, a batch re-balancing plan is put in place to facilitate improved batch distribution. The CNN model prototype has been created and evaluated with the TensorFlow2 backend in the Keras profound learning library. The graphical enactment estimation of qualified profound DL classifiers with loss and accuracy in the testing and training phase. The finest results were achieved in training and testing accuracy. The resulting confusion matrices in the tested deep learning classification are shown in Fig. 5.

To check the classification efficiency of the DL classifier with the false positive rate (FPR), true positive rate (TPR) we have also added a receiver operating characteristic (ROC) curve to distinguish positive nCoV-19 cases in CXR images tested as shown in Fig. 6.

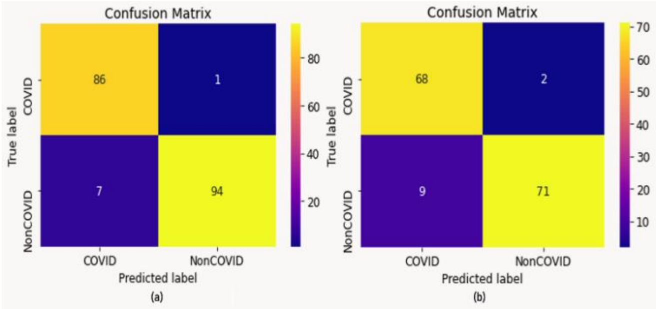


Fig. 5. Confusion Matrix of the Model with different Batch Sizes (a) 16 Batch Size (b) 32 Batch Size.

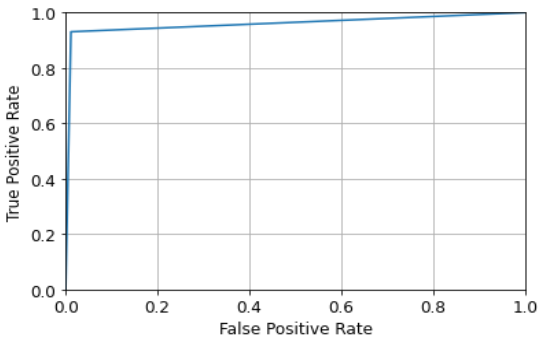


Fig. 6. Receiver Operating Characteristic (ROC) Curve.

A. Dataset and Experimental Setup

Several publicly accessible data-sets provide a huge number of CXR images. For instance, nCoV-19 is very innovative, none of the broad repositories contains labeled nCoV-19 data, so we depend on a minimum of two data-sets for nCoV-19 and normal images. The publicly open COVID-19 Image Data Collection [49] collected nCoV-19 chest x-rays. The images are predictably variable in size and quality from this series. In this dataset collection, there were a total of 115 PA images classified as COVID-19. The images are very different in resolution with a smaller pixel scale of 224×224 and a larger pixel of 1024×1024. Contrast, brightness, and subject locating are all extremely variable in this data-set. The data collection primarily contains adult patient data. The image samples used for this research are shown in Fig. 7 where Fig. 7(a) represents Positive nCoV-19 CXR and Fig. 7(b) represents Negative nCoV-19 CXR. National Institute of Health (NIH) CXR [50] foundation, 112, 120 anonymized X-Ray imaging with 14 condition labels including pneumonia and normal conditions. The images of the COVID-19 data-set have a similar quality, size, and feature ratio to the regular images with dimensions typically 1024×1024 pixels of portrait orientation. This data-set was chosen as the source of pneumonia and regular X-Ray images. Table I summarizes our findings concerning the data sources available:

In this analysis, we aim to use real chest x-ray data and not at this point to create and use synthetic data. For our model experiments, we also intended to use a balanced data-set size. The master data-set containing COVID-19, Pneumonia, and normal Images for our model creation and testing purposes has been collected from the two source data-sets COVID-19 and NIH. The COVID-19 data-set was designed to eliminate images that were the incorrect projection, low resolution, and unwell cropped. We had more than 100 usable samples for a data-set of COVID-19. Since we use the NIH data-set to pick appropriate numbers of samples for normal and infected cases, we use sampling methods to select several more samples. The Pneumonia and Normal sample images from the downloaded NIH data-set randomly selected exclude samples for young patients. To draw the attention of machine learning (ML) algorithms, a few images so selected contained medical devices which we were concerned about so that these images were discarded and replaced by a random selection until those devices were absent from the data-set. Table II summarizes the outcomes of this process. The radio-graphs showed a Ground Glass Opacity/Opacification (GGO) pattern for reported COVID-19 disease, with rarely combined patches, peripherals, and bilateral. The methodology was established using a software bundle (Anaconda3). The execution was GPU precise. All trials were done on a Dell Inspiron 5570 Core(TM) i5-8250U CPU at 1.6 GHz (8 CPUs), 16 GB of RAM. All experiment trials were carried out with 70% of the data-set for training while 25% of the dataset for testing the leftover 5% is for validation.

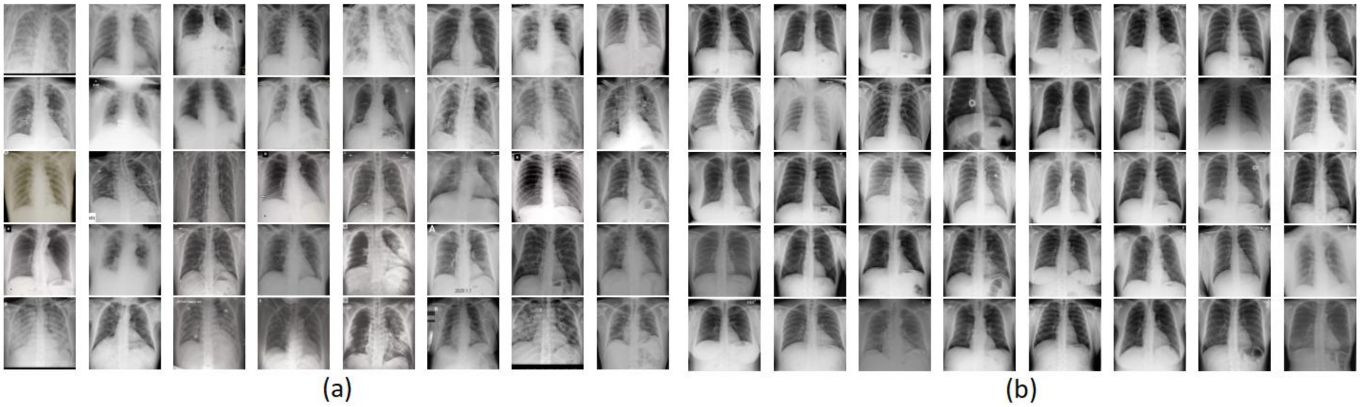


Fig. 7. Sample of Images used in Research (a) Positive nCoV-19 Images (b) Negative nCoV-19 Images.

TABLE I. SUMMARY OF DATA SOURCES CONSIDERED

Collection	No. of Images	Characteristics
COVID-19 Image Data Collection [44]	115 COVID-19 (PA)	Flexible size, feature etc.
NIH Chest X-Ray [45] No Conclusion	322 Pneumonia, 60360	Images are 1024×1024 pixels.

TABLE II. SAMPLED DATA-SET FOR THE EXPERIMENT

Source	Condition	No. of source images	No. of curated images
NIH CXR	Normal	60361	200
NIH CXR	Pneumonia	322	322
nCoV-19 Image Data Collection	COVID-19	115	100

B. Testing Accuracy and Confusion Matrix

Testing accuracy is an estimate which shows the accuracy and precision of any selected deep model. Besides, the Confusion Matrix (CM) is a quantitative metric that offers further information about the accuracy of the test achieved. The confusion matrix of the model presented in Fig. 5(a) shows CM with batch size 16 and Fig. 5(b) shows CM with batch size 32. The result of the training and testing model is shown below Fig. 8(a) shows the model with batch size 16 and Fig. 8(b) shows the model with batch size 32. The accuracy of the model was calculated utilizing the equation (1). For 16 and 32 batch sizes the model gives an accuracy of 95.74% and 92.66% respectively. The succeeding output metrics can be determined after computing the ideals batch size 16 possible consequences in the CM.

1) Accuracy: The significant metric for the outcomes of DL classifiers, as specified in equation (1). It is the ∑ TP and TN divided by the total values of the CM.

$$\text{Accuracy} = \frac{TP+TN}{\text{TP+FP+TN+FN}} \quad (1)$$

2) Precision: Denoted in equation (2) to provide the association between the TP predicted values and total positive predicted values.

$$\text{Precision} = \frac{TP}{\text{TP+FP}} \quad (2)$$

3) Recall: In equation (3), Recall is the fraction among the TP values of estimate and the ∑ expected TP values and expected FN values.

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

4) F1-score: F1-score is a total degree of the accurateness that chains the precision and recall, as denoted in equation (4). F1-score is the double of the fraction among the multiplication to the ∑ of recall and precision metrics.

$$\text{F1} - \text{Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

5) Sensitivity: Measures the fraction of actual positives that are properly recognized, as represented in equation (5) as such (e.g., the ratio of sick individuals who are properly recognized as having the disorder).

$$\text{Sensitivity} = \frac{TP}{\text{TP+FN}} \quad (5)$$

6) Specificity: (Also termed as the TN rate) trials the fraction of concrete negatives that are properly recognized, as represented in equation (6) as such (e.g., the ratio of fit individuals who are properly recognized as not having the disorder).

$$\text{Specificity} = \frac{TN}{\text{TN+FP}} \quad (6)$$

One of the benefits of generating a confusion matrix is measuring the accuracy of the tests. Precision, Recall, F1 Score, Sensitivity, and Specificity are provided in the following equations (2-6). They are the most important performance measurements in the field of DL. Table III provides information regarding the enactment of the model, respectively in terms of specificity, sensitivity, precision, and F1 scores.

TABLE III. PERFORMANCE OF THE MODEL

Batch Size	Accuracy	Specificity	Sensitivity	Precision	F1-Score
16	95.74%	98.94%	92.47%	98.85%	95.55
32	92.66%	97.26%	88.31%	97.14%	92.60

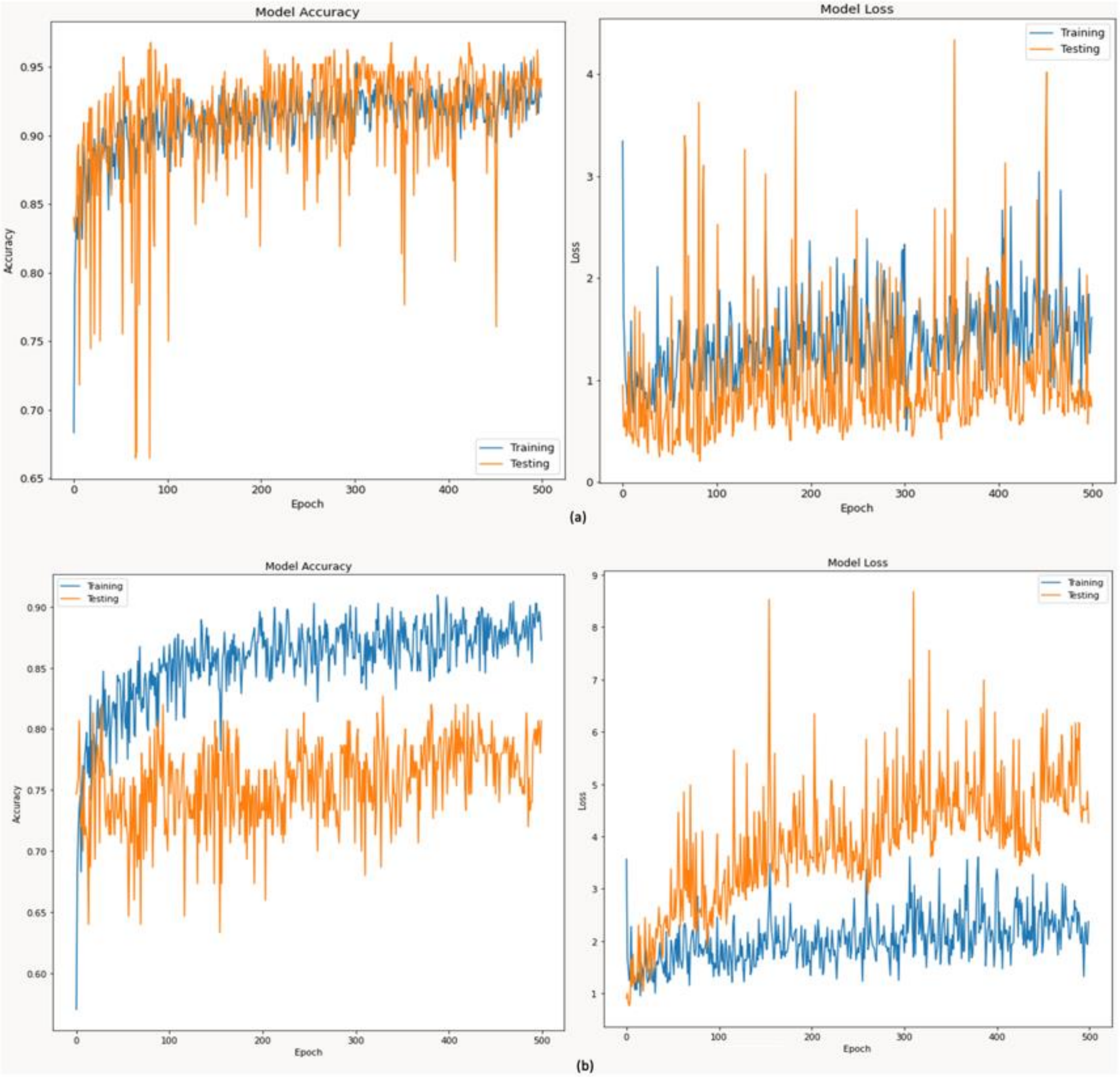


Fig. 8. Model Accuracy and Model Loss with different Batch Sizes (a) 16 Batch Size (b) 32 Batch Size.

VI. Discussion

The number of people in over one hundred and seventy countries has been infected by the COVID-19 up to now, and this number may, unfortunately, increase over the coming days. COVID-19 has been declared a pandemic by the WHO. The use of the deep learning system approached will show promising results for identification by digital images of CXR of morphological changes in the lungs of COVID-19 infected patients [51]. In this analysis, InceptionV3 was used with an accuracy of 95.74% and the F1 score is 95.55. One of the limitations of this analysis is the number of cases. The organization gathering COVID-19 patient data from various countries of the world can be strategic and investigated to further improve the investigative method. With the variation of some parameters, different results are achieved which are shown in Table III. The interpretations accomplished from the present research verified that the proposed approach is capable and can be further applied to the multi-step on the prediction of a diverse set of disease parameters such as Cancer, Cardiac, Infectious, and Liver Disease. In the upcoming, we plan to authenticate our model by integrating more CXR images. This should decrease clinician capacity expressively. We will address CT images for nCoV-19 recognition and associate the gained outcomes using the suggested model trained to utilize CXR images. Also, we will attempt to assemble more native radiology images for nCoV-19 cases and estimate them using our CNN model from locations in Pakistan. Afterward, the compulsory trials are finished, we target to set up the established model in native hospitals for screening.

VII. Conclusion

In this research, we presented a Deep Convolutional Neural Network (DCNN) strategy for the recognition of nCoV-19 cases from CXR images that is open-source and accessible to the broad community, which could be used as an auxiliary tool for extremely guarded health specialists in determining the sequence of treatment. We also defined a hybrid CXR data-set to train DCNN that is contained 10,000 CXR images from two open access data repositories, also we have added 150 new cases of the COVID-19^+ CXR images. Additionally, we examined how the established model makes predictions in a shot to increase deeper insights into critical aspects related to COVID-19 cases, which can support clinicians in enhanced screening as well as transparency when DCNN for enhanced computer-aided screening (CAS). Our outcomes of around 95% for both precision and sensitivity are realistic initially, moreover that proficient clinicians have an explanatory radiological error of 2-10% liable on the radiological examination. We set up that the InceptionV3 classifier provided decent results within the tentative limitations of the small figures of presently available nCoV-19 chest x-rays. As the figures of existing nCoV-19 CXR images rise, we'll be capable to offer sufficient training data quantity to the deeper network and thus obtain superior outcomes from the InceptionV3 classifier. We aim to make our CNN model further reliable and correct utilizing extra chest x-ray images from our native hospitals.

Acknowledgment

We would like to thank the Mehran University of Engineering and Technology, Jamshoro, Pakistan, for providing us the necessary assets to conduct this research.

References

[1] X. Xu et al., "Deep Learning System to Screen Coronavirus Disease 2019 Pneumonia," pp. 1–29, 2019.

[2] N. E. M. Khalifa and M. H. N. Taha, "Detection of Coronavirus (COVID-19) Associated Pneumonia based on Generative Adversarial Networks and a Fine-Tuned Deep Transfer Learning Model using Chest X-ray Dataset," pp. 1–15.

[3] M. Loey, F. Smarandache, and N. E. M. Khalifa, "Within the Lack of COVID-19 Benchmark Dataset: A Novel GAN with Deep Transfer Learning for Corona-virus Detection in Chest X-ray Images," no. April 2020.

[4] W. H. Organization, "Naming the coronavirus disease (COVID-19) and the virus that causes it," URL https//www. who. int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it, 2020.

[5] A. Narin, C. Kaya, and Z. Pamuk, "Department of Biomedical Engineering, Zonguldak Bulent Ecevit University, 67100, Zonguldak, Turkey."

[6] M. Ilyas, H. Rehman, and A. Nait-ali, "Detection of Covid-19 From Chest X-ray Images Using Artificial Intelligence: An Early Review," no. April 2020, [Online]. Available: http://arxiv.org/abs/2004.05436.

[7] https://bluedot.global/ [last visited 07/04/2020].

[8] I. I. Bogoch, A. Watts, A. Thomas-Bachli, C. Huber, M. U. G. Kraemer, and K. Khan, "Pneumonia of unknown aetiology in Wuhan, China: Potential for international spread via commercial air travel," J. Travel Med., vol. 27, no. 2, pp. 1–3, 2020, doi: 10.1093/jtm/taaa008.

[9] http://www.diseasedaily.org/about [last visited 07/04/2020].

[10] H. S. Maghded, K. Z. Ghafoor, A. S. Sadiq, K. Curran, D. B. Rawat, and K. Rabie, "A Novel AI-enabled Framework to Diagnose Coronavirus COVID-19 using Smartphone Embedded Sensors: Design Study," Proc. - 2020 IEEE 21st Int. Conf. Inf. Reuse Integr. Data Sci. IRI 2020, pp. 180–187, 2020, doi: 10.1109/IRI49571.2020.00033.

[11] M. Akhtar, M. U. G. Kraemer, and L. M. Gardner, "A dynamic neural network model for predicting risk of Zika in real-time," bioRxiv, pp. 1–17, 2018, doi: 10.1101/466581.

[12] https://www.technologyreview.com/s/615360/cdc-cmu-forecasts-coronavirus-spread/ [last visited 07/04/2020].

[13] M. A. Zaki and U. Zaki, "CHALLENGES OF ENGLISH TEXT RECOGNITION FROM NATURAL SCENES," Glob. Sci. J., vol. 8, no. 4, pp. 1045–1054, 2020, [Online]. Available: http://www.globalscientificjournal.com/researchpaper/CHALLENGES_OF_ENGLISH_TEXT_RECOGNITION_FROM_NATURAL_SCENES.pdf.

[14] U. ZAKI et al., "Dataset of Urdu ud1k from Natural Scenes," SindhUniv. Res. Jour. (Sci. Ser.), vol. 51, no. 04, pp. 595–600, 2019.

[15] https://deepmind.com/research/open-source/computational-predictions-of-protein-structures-associated-withCOVID-19 [last visited 07/04/2020].

[16] [16] https://www.healthmap.org/covid-19/?mod=article-inline [last visited 07/04/2020].

[17] https://nextstrain.org/ncov [last visited 07/04/2020]

[18] https://www.againstcovid19.com/singapore/dashboard [last visited 07/04/2020].

[19] https://www.scmp.com/comment/opinion/article/3075553/time-coronavirus-chinas-investment-ai-paying-bigway [last visited 07/04/2020].

[20] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network," pp. 1–9, 2020.

[21] H. S. Maghdid, A. T. Asaad, K. Z. Ghafoor, A. S. Sadiq, and M. K. Khan, "Diagnosing COVID-19 Pneumonia from X-Ray and CT Images using Deep Learning and Transfer Learning Algorithms," pp. 1–8.

[22] L. O. Hall, R. Paul, D. B. Goldgof, and G. M. Goldgof, "Finding COVID-19 from Chest X-rays using Deep Learning on a Small Dataset," pp. 1–8.

[23] X. Li, C. Li, and D. Zhu, "COVID-MOBILEXPERT: ON-DEVICE COVID-19 SCREENING USING SNAPSHOTS OF CHEST X-RAY," 2020.

[24] S. Rajaraman, J. Siegelman, P. O. Alderson, L. S. Folio, R. Les, and S. K. Antani, "Iteratively Pruned Deep Learning Ensembles for COVID-19 Detection in Chest X-rays," pp. 1–10.

[25] S. Basu, S. Mitra, and N. Saha, "Deep Learning for Screening COVID-19 using Chest X-Ray Images," pp. 1–6, 2020.

[26] S. Narejo and E. Pasero, "An application of internet traffic prediction with deep neural network," Smart Innov. Syst. Technol., vol. 69, pp. 139–149, 2017, doi: 10.1007/978-3-319-56904-8_14.

[27] S. Narejo, E. Pasero, and F. Kulsoom, "EEG based eye state classification using deep belief network and stacked autoencoder," Int. J. Electr. Comput. Eng., vol. 6, no. 6, pp. 3131–3141, 2016, doi: 10.11591/ijece. v6i6.12967.

[28] S. C. Cheng et al., "First case of Coronavirus Disease 2019 (COVID-19) pneumonia in Taiwan," J. Formos. Med. Assoc., vol. 119, no. 3, pp. 747–751, 2020.

[29] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2014.

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 vol. 2017-January, ed: Institute of Electrical and Electronics Engineers Inc., pp. 2261-2269, 2017.

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818-2826, 2016.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE Computer Society Conference on

Computer Vision and Pattern Recognition, vol. 2016-December, pp. 770-778, 2016.

[33] E. C. Too, L. Yujian, S. Njuki, and L. Yingchun, "A comparative study of fine-tuning deep learning models for plant disease identification," Computers and Electronics in Agriculture, vol. 161, pp. 272279, 2019/06/01/ 2019.

[34] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," presented at the Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, California, USA, 2017.

[35] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 vol. 2017-January, ed: Institute of Electrical and Electronics Engineers Inc., 2017, pp. 1800-1807.

[36] M. A. ZAKI, S. ZAI, M. AHSAN, and U. ZAKI, "DEVELOPMENT OF AN ANDROID APP FOR TEXT DETECTION," J. Theor. Appl. Inf. Technol., vol. 97, no. 20, pp. 2485–2496, 2019.

[37] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 4510-4520, 2018.

[38] P. K. Sethy and S. K. Behera, "Detection of coronavirus Disease (COVID-19) based on Deep Features," Https://Www.Preprints.Org/Manuscript/202003.0300/V1, no. March, p. 9, 2020.

[39] F. Shan+ et al., "Lung infection quantification of covid-19 in ct images with deep learning," arXiv Prepr. arXiv2003.04655, 2020.

[40] C. Butt, J. Gill, D. Chun, and B. A. Babu, "Deep learning system to screen coronavirus disease 2019 pneumonia," Appl. Intell., p. 1, 2020.

[41] S. Wang et al., "A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19)," MedRxiv, 2020.

[42] A. Bhandary et al., "Deep-learning framework to detect lung abnormality–A study with chest X-Ray and lung CT scan images," Pattern Recognit. Lett., vol. 129, pp. 271–278, 2020.

[43] O. Stephen, M. Sain, U. J. Maduh, and D.-U. Jeong, "An efficient deep learning approach to pneumonia classification in healthcare," J. Healthc. Eng., vol. 2019, 2019.

[44] E. Ayan and H. M. Ünver, "Diagnosis of Pneumonia from Chest X-Ray Images Using Deep Learning," in 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), 2019, pp. 1–5.

[45] D. Varshni, K. Thakral, L. Agarwal, R. Nijhawan, and A. Mittal, "Pneumonia Detection Using CNN based Feature Extraction," in 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019, pp. 1–7.

[46] M.Z. Alom, T.M. Taha, C. Yakopcic, S. Westberg, SM. Hasan, B.C. Van Essen, A.A.S. Awwal and V.K. Asari, The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches. arXiv preprintarXiv:1803.01164, 2018.

[47] G. Litjens et al., "A survey on deep learning in medical image analysis," Med. Image Anal., vol. 42, pp. 60–88, 2017, doi: https://doi.org/10.1016/j.media.2017.07.005.

[48] M. Rahimzadeh and A. Attar, "A NEW MODIFIED DEEP CONVOLUTIONAL NEURAL NETWORK FOR DETECTING COVID-19 FROM X-RAY IMAGES," 2020.

[49] J.Cohen, P. Morrison and L. Dao, "COVID-19 Image Data Collection" arXiv.org, 2020.

[50] "National Institutes of Health - NIH Clinical Center provides one of the largest publicly available chest x-ray datasets to the scientific community," ed: Normans Media Ltd., 2017.

[51] J. Melendez et al., "An automated tuberculosis screening strategy combining X-ray-based computer-aided detection and clinical information," Sci. Rep., vol. 6, p. 25265, 2016.

Secure Software Engineering: A Knowledge Modeling based Approach for Inferring Association between Source Code and Design Artifacts

Chaman Wijesiriwardana[1]
Faculty of Information Technology
University of Moratuwa
Katubedda, Sri Lanka

Ashanthi Abeyratne[2], Chamal Samarage[3],
Buddika Dahanayake[4], Prasad Wimalaratne[5]
University of Colombo School of Computing
Reid Avenue, Colombo 07, Sri Lanka

***Abstract*—Secure software engineering has emerged in recent decades by encouraging the idea of software security has to be an integral part of all the phases of the software development lifecycle. As a result, each phase of the lifecycle is associated with security-specific best practices such as threat modeling and static code analysis. It was observed that various artifacts (i.e., security requirements, architectural flaws, bug reports, security test cases) generated as a result of security best practices tend to be segregated. This creates a significant barrier to resolve the security issues at the implementation phase since most of them are originated in the design phase. In order to address this issue, this paper presents a knowledge-modeling based approach to semantically infer the associations between architectural level security flaws and code-level security bugs, which is manually tedious. Threat modeling and static analysis are used to identify security flaws and security bugs, respectively. The case study based experimental results revealed that the architectural level security flaws have a significant impact on originating security bugs in the code level. Besides, the evaluation results confirmed the scalability of the proposed approach to large-scale industrial software products.**

***Keywords*—Software security; threat modeling; knowledge modeling; security flaws**

I. Introduction

Having identified the critical need for software security, the paradigm shift of *"Building Security In"* has emerged in the recent decades [1], [2], [3]. This paradigm shift requires software security to be addressed in all phases of the software development lifecycle. Literature reveals that most security vulnerabilities result from defects that are unintentionally introduced in the software during the design phase and the implementation phase [2]. Garry McGraw has identified code reviews and architectural risk analysis as the top two best practices to minimize the security vulnerabilities in software systems [2]. These best practices are called as *security touchpoints* associated with the artifacts produced by the implementation phase (i.e., codebase) and the design phase (i.e., design documents) respectively. Even in organizations with mature software development processes, the artifacts created are segregated from each other [4]. Furthermore, to the best of our knowledge, existing tools are not capable of identifying security-specific associations between the artifacts generated during software development. This reveals a significant research gap of interlinking the artifacts originated at the implementation phase and the design phase.

This paper presents a conceptual framework and a proof-of-concept implementation to semantically interlink architectural level security flaws and code-level security bugs based on the foundation laid in [5]. Security flaws are identified based on STRIDE [6], [7] threat categorization model introduced by Microsoft, which helps to identify threats from the attackers' perspective by classifying attackers goals into six threat categories. Security bugs are determined based on OWASP Top 10 [8], [9], [10] vulnerabilities, which are the ten most critical web application security risks providing a great awareness for web application security. In this paper, security flaws and bugs are interlinked by employing a knowledge-modeling based technique, which facilitates inferring the associations that are manually tedious.

In this approach, rather than directly interlinking different artifacts, it is required to infer the associations among design documents and source code to reveal whether the root causes for security bugs lie in the design phase. Knowledge-modeling based approaches are capable of handling large quantities of data intelligently [11] and proven successful in the domain of cybersecurity [12]. Besides, expertise in software security is not readily available. Therefore, knowledge-modeling approaches are useful when security expertise is not available. It also provides a common platform for integrating knowledge on a large scale. Finally, knowledge bases are capable of generating new knowledge by using the stored data.

More precisely, the key contributions of this paper are:

- a knowledge model and the association rules to infer the hidden relationships across security flaws and bugs, and
- results of the evaluation experiments conducted by using the proof-of-concept implementation.

The remainder of this paper is organized as follows: Section 2 presents the related work. Section 3 describes the proposed approach to interlink security artifacts. Section 4 presents the proof-of-concept implementation followed by the evaluation in Section 5. Section 6 concludes and present future work.

II. Background and Related Work

Building Security In[2] is a collaborative effort that provides practices, tools, guidelines, rules, principles, and other resources that software developers, architects, and security practitioners can use to build security into software in every phase of its development. Correspondingly Microsoft has carried out a noteworthy effort under its Trustworthy Computing Initiative which focused on people, process, and technology to tackle the software security problem [13]. On the people front, Microsoft trains every developer, tester, and program manager in basic techniques of building secure products. Microsoft's development process has been enhanced to make security a critical factor in design, coding, and testing of every product. A key part of Microsoft's Trustworthy Computing is the Security Development Lifecycle (SDL) which focuses on software development and introduces security and privacy throughout all phases of the software development process. The Microsoft SDL combines a holistic and practical approach to reduce the number and severity of vulnerabilities in Microsoft products [1]. Conforming to the aforementioned approaches introduced to the SDLC, it conveys that Architectural risk analysis and Code review are two significant steps which should be conducted in a security specific SDLC process.

A. Architectural Risk Analysis

Frydman et al. [14] introduce an automated approach for threat modeling by producing two data structures: identification trees and mitigation trees. Identification trees are used to determine threats in the software design, while mitigation trees describe countermeasures of threats by classifying software specifications that are required to mitigate a specific risk. The two data structures and ranking information of threats have combined in a knowledge base called *attack patterns*. Yuan et al. [15] describe their approach to develop a tool to retrieve relevant common attack pattern enumeration and classification (CAPEC) type attack patterns for software development. CAPEC attack patterns are valuable resources that can help software developers to think like an attacker and have the potential to be used in each phase of the secure software development lifecycle. A metric has been defined in this tool to measure the degree of usefulness of an attack pattern and the degree of its relevance to a particular STRIDE category. Berger et al. [16] propose a practical approach to architectural risk analysis that leverages Microsoft threat modeling approach. This proposed approach uses extended DFDs and a security knowledge base to aid software developers in detecting vulnerabilities in software architectures. The knowledge base contains information on architectural weaknesses and possible mitigations. However, these approaches explicitly operate only on the design phase with no effort to interlink the threats with source code level bugs.

B. Security Specific Code Analysis

A practical approach for implementing secure practices into the software development lifecycle outlined in [17]. It has introduced a development testing platform which allows the development organizations to coherently integrate code testing into the software development process. Coverity development testing solutions train developers to address both security and quality when testing the code which leads to secure software development practices. The commonly found potentially critical security defects in the source code are identified from this platform, which is an aid for the developers to fix them. The major weakness of this platform is that it mainly focuses on implementation phase without considering the actual root-causes lie in the design phase. Alqahtani et al. [18] have stressed the fact that despite the security concerns are reported in specialized vulnerability databases; these repositories often remain information silos. As a solution, they introduced a modeling approach that eliminates these silos by linking security knowledge with other software artifacts to improve traceability and trust in software products. A Security Vulnerabilities Analysis Framework (SV-AF) is introduced in this approach to support evidence-based vulnerability detection. This approach also explicitly focus the code level and connecting the design phase with the identified bugs is not addressed.

1) Interlinking static analysis with other artifacts:

Interlining of software artifacts has been extensively discussed in the literature [19], [20], [21]. Implementation vulnerabilities differentiate themselves from the design vulnerabilities because they only exist in the source code and are not part of the original design or requirements. The implementation vulnerabilities are also very language-specific, especially the C and C++ coding languages are infamous for their ease of creating implementation vulnerabilities [22], [23]. The languages memory control is both its strength and weakness. The control the developer has creates the opportunity to create optimized and fast software but also insecure code that can easily be exploited. Some of the most common causes of implementation vulnerabilities are buffer overflows, format string bugs, integer overflows, null dereferences, and race conditions.

A novel approach for implementing secure practices into the software development lifecycle outlined in [17]. It has introduced a development testing platform which allows the development organizations to coherently integrate code testing into the software development process. Coverity development testing solutions train developers to address both security and quality when testing the code which leads to secure software development practices. The commonly found potentially critical security defects in the source code are identified from this platform, which is an aid for the developers to fix them. The major weakness of this platform is that it mainly focuses on implementation phase without considering the actual root-causes lie in the design phase. Alqahtani et al. [18] have stressed the fact that despite the security concerns are reported in specialized vulnerability databases; these repositories often remain information silos. As a solution, they introduced a modeling approach that eliminates these silos by linking security knowledge with other software artifacts to improve traceability and trust in software products. A Security Vulnerabilities Analysis Framework (SV-AF) is introduced in this approach to support evidence-based vulnerability detection. This approach also explicitly focus the code level and connecting the design phase with the identified bugs is not addressed.

In contrast, our approach is not limited to ensuring the security in a single phase of the software development lifecycle or a single artifact originated from a specific lifecycle phase. Instead, we attempt to semantically interlink the artifacts produced in the design phase and implementation phase. It allows software practitioners to identify the root-causes for the security bugs.

III. Approach

This approach aims at inferring the associations between security flaws and security bugs introduced during the design phase and the implementation phase of the lifecycle. As stated previously, security flaws are identified in terms of STRIDE threat categorization, and security bugs are represented regarding OWASP top 10 vulnerabilities. The approach consists of three main constituents: *threat modeling* to identify security flaws, *static code analysis* to identify security bugs, and exploiting *knowledge base* to infer relationships among flaws and bugs. The conceptual architecture of the proposed approach, exhibiting its external data sources, internal components, boundaries, and their associations, is depicted in Fig. 1.

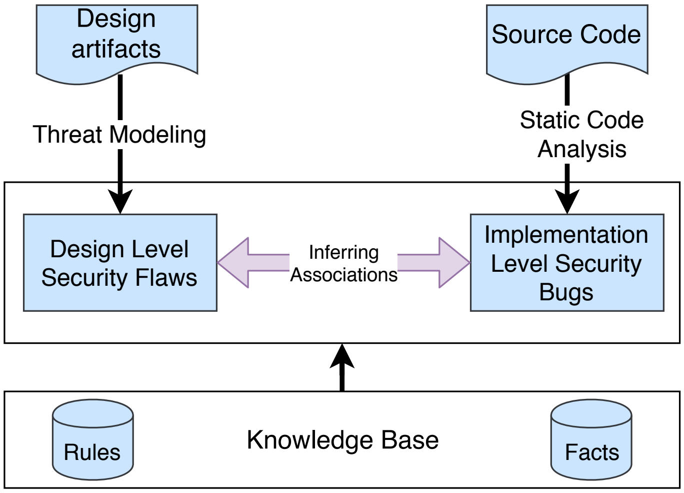


Fig. 1. Conceptual Model to Infer Associations between Flaws and Bugs.

A. Identifying and Pre-Processing Security Flaws

Security flaws are identified through an architectural risk analysis, which includes explicitly identifying security risks in the software architecture/design. In this paper, threat modeling used as the architectural risk analysis method due to several noteworthy reasons such as the ability to work with high-level design diagrams, simplicity to employ in different contexts, and explicit tool support. For example, threat modeling process can be initiated by drawing a data flow diagram (DFD). According to Abi-Antoun et al. [24], architectural level security flaws can effectively identify by analyzing Level 0 or Level 1 DFDs. As depicted in Fig. 2, threat modeling process consists of thee steps.

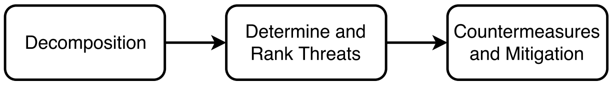


Fig. 2. Main Steps of the Threat Modeling Approach.

1) Decomposition:: This step concerned with gaining an understanding of the application and how it interacts with external entities. This knowledge helps in identifying entry points to see where a potential attacker could interact with the application, determining trust levels which represent the access rights that the application will grant to external entities and identify assets that the attacker would be interested. Thus, this information is used to produce data flow diagrams (DFDs) for the application.

2) Determine and rank threats:: In this step, threats are determined and categorized according to a threat categorization methodology. The goal of threat categorization is to identify threats from both attackers perspective and defensive perspective. DFDs produced in step 1 is primarily used to identify potential threat targets from the attackers perspective.

3) Countermeasures and mitigation:: In this step, mitigations, and countermeasures are identified for the ranked threats.

B. Using Static Analysis to Identify Security Bugs

Static analysis is used to detect the security bugs that appear in the source code of the software project. For adequately understanding the code level security bugs, they are categorized based on OWASP Top 10 vulnerabilities. OWASP Top 10 is the ten most critical web application security risks which provide a powerful awareness document for web application security. The different versions of OWASP T10 are focused on identifying the most common vulnerabilities which depict how an attacker can potentially harm a software system. Though the static analysis detects the bugs that are categorized into OWASP vulnerabilities, that information is not sufficient to generate a relationship with security flaws. Therefore, it was decided to utilize OWASP Proactive Controls[1], which is a set of developer-centric security techniques that can be included in the every software project. Most importantly, each proactive control helps in preventing one or more of the OWASP Top Ten web application security vulnerabilities. OWASP top 10 vulnerabilities have been mapped to proactive controls.

C. Inferring Relationships among Flaws and Bugs

Inferring logical relationships between security flaws and bugs are expected to obtain via STRIDE and OWASP. However, STRIDE mainly focuses on attacking perspective of software security. On the other hand, Application Security Frame (ASF)[2] is a threat categorization model, which helps to identify the threats from the defensive perspective. For an in-depth analysis of the threats affecting the software application data and functional assets, both the STRIDE attacker view and the ASF defensive view for the enumeration of threats considered as essential. As stated previously, threat modeling process only focuses on attacker's perspective based on STRIDE. Therefore, to involve the defensive standpoint, a relationship has been identified between STRIDE and ASF.

What the relationship depicts by UcedaVelez and Morana [25] is not a complete association between ASF and STRIDE due to each category of STRIDE lacks an association to ASF type. Hence, this association further improved by combining the findings of Adam Shostack[6]. Table I presents the improved mapping between STRIDE and ASF.

[1] https://www.owasp.org
[2] https://msdn.microsoft.com/en-us/library/ff649461.aspx

TABLE I. Mapping between STRIDE and ASF

STRIDE Attack Type	ASF type
Spoofing	Authentication
Tampering Information Disclosure Elevation of privileges	Authorization
Repudiation Elevation of privileges	Configuration Management
Tampering Information Disclosure	Data Protection in Storage and Transit
Tampering	Data Validation / Parameter Validation
Information Disclosure	Error Handling and Exception Management

1) Semantic Text Similarity between ASF and Proactive Controls: The semantic text similarity calculated for every single security control in ASF with every single Proactive Control. The descriptions of ASF security controls and Proactive Controls are not limited to a single phrase. Accordingly, the semantic text similarity of each phrase of the description of a particular ASF security control calculated concerning each phrase of the description of Proactive Control. Consequently, the average semantic similarity score between a specific ASF security control (Ai) and Proactive control (Pi) calculated as follows.

$$SemS(Ai, Pi) = \left[\left(\sum_{i=0}^{nm} Vi \right) \div nm \right] \tag{1}$$

where:

A_i = description of ASF having n phrases
P_i = description of proactive controls having m phrases
V_i = similarity between ASF and proactive control

2) Knowledge Base: Security specific information about software projects can be in in the form of either structured or unstructured in heterogeneous information sources. Some of these information sources may frequently undergo significant changes as well. Thus, a knowledge modeling approach would more practical and beneficial in developing a security framework for software development. The knowledge base of this approach contains the facts and rules related to the STRIDE, ASF, OWASP T10, Proactive Controls and Semantic Similarity Scores between ASF and Proactive controls. A Frame-based approach is used for knowledge representation of facts [26]. The structure of the frames for the facts STRIDE, OWASP T10, and Similarity Matching are as follows.

Listing 1: Frame for STRIDE Categories

```
frame (stride,
[category_model [val threat],
types - [val [spoofing, tampering,
information disclosure,
denial of service, elevation of privileges]]
])
```

Listing 2: Frame for OWASP Categories

```
frame (owasp_t10,
[category_model [val bug],
types - [val [a1, a2, a3, a4, a5,
                  a6, a7, a8, a9, a10]]
```

Listing 3: Frame for Semantic Similarity

```
frame(semantic_similarity
[proactive_control [val c1],
security_control [val s1],
score [val value]
```

Prolog rules were designed to infer the association between STRIDE and OWASP T10.

Rule 1: Querying the Knowledge-base

```
isCausedByThreatCategories
(BugCategory, TList_Unique) :-
findall(T, isCausedByThreatCategory
(BugCategory, T), TList),
sort(TList, TList_Unique)
```

Rule 1 is used to query the knowledge base. The list of unique threat categories can be discovered by querying the knowledge base using a bug category. Each threat category associated with bug category is revealed by the Rule 2.

Rule 2: Discovering the associated threat category using the bug category

```
isCausedByThreatCategory(BugCategory, T) :-
lacksProactive(BugCategory, P),
mapsToSecurityControl(P, S),
isWeakendByThreatCategory(S, T)
```

Rule 2 is used to discover the associated threat category using the bug category. The threat category is revealed using the subsequent rules on the right-hand side of Rule 2. The lacksProactive(BugCateogry, ProactiveControl) is used to discover the proactive controls violated due to the given bug category.

Rule 3 - Identifying the proactive controls of the relevant bug categories

```
lacksProactive(BugCategory, C) :-
isProactiveListOf(CList, BugCategory),
member(C, CList)
```

Rule 3 is used to identify the proactive controls of the relevant bug categories in succession.

Rule 4 - Discovering the associated threat category using the bug category

```
isProactiveListOf(CList, BugCategory) :-
owasp_top10(BugCategory, _ , CList)
```

The Rule 4, isProactiveListOf(ProactiveControlList, BugCategory) used to identify the proactive list of the given bug category using the owasp_top10 frame.

Rule 5 - Discovering the associated threat category using the bug category

```
mapsToSecurityControl(Proactive, S) :-
isMappingSecurityControlList(SList, Proactive),
member(S, SList)
```

Rule 5 is used to identify the mapping ASF security controls in succession by using the semantic text similarity score between ASF security controls and proactive controls. An ASF security control is mapping with a proactive control if it belongs to the top three semantic text similarity scores of the relevant proactive control. Additional six rules are used to identify the mapping security controls using the semantic text similarity scores, which is not listed here due to space limitations.

Rule 6 - Discovering the associated threat category using the bug category

```
isWeakendByThreatCategory(SecurityControl,T):-
stride(_, T, _, SecContList),
member(SecurityControl, SecContList)
```

The association results given by the Knowledge-base are used to create the associations between Bugs and Threats. The facts regarding STRIDE and ASF security controls are static facts while OWASP T10 and Proactive Controls are dynamic facts. The reason for keeping OWASP T10 and Proactive Controls as dynamic is that both of these facts are continuously getting revised based on technological advancements and industrial best practices. Thus, the knowledge base generates new knowledge based on the regular updates.

IV. Proof-of-Concept Implementation

In this research, as a proof-of-concept, a security analysis framework has implemented to infer the relationships between design flaw and implementation bugs by adhering to the theoretical foundation described in the previous section. Figure 3 provides an overview of the proof-of-concept implementation and its main constituents: (a) STRIDE Categorization of Security Flaws, (b) OWASP Categorization of Security Bugs, and (c) Knowledge-base and Association inferencing.

A. STRIDE Categorization of Security Flaws

As described previously, design-level security flaws are identified by using Threat Modeling. A threat model will be generated by analyzing the Level-0 or Level-1 DFD of the software system under investigation. The threat model is created using the Microsoft Threat Modeling Tool (TMT), which categorizes the threats according to the STRIDE model. The threat model generated from Microsoft TMT obtained as an XML file, which is further processed to extract the *threats*. After that, the extracted threats converted into *threat objects* that contains the relevant details of the threats introduced based on the proposed design as depicted in the DFD. Then *threat category objects* are created based on STRIDE categorization, which facilitates identifying the specific threat category of a specific threat introduced in the design phase. Users can either upload an external DFD or manually draw it.

B. OWASP Categorization of Security Bugs

SA-SEC facilitates code analysis in two distinct ways. First, it can use SonarQube to analyze the source code. SonarQube allows the categorization of the vulnerabilities identified as security bugs into OWASP Top 10. However, *SA-SEC* does not entirely depend on a third-party tool like Sonarqube for bug categorization. A novel mechanism based on Case-based reasoning [27] is introduced to categorize the bugs into OWASP categories by analyzing the different attributes of identified vulnerabilities. A collection of attributes such as *threat description, threat type, etc* that are common across multiple security tools have identified. Table II provides a list of the selected attributes together with a short description.

TABLE II. Attributes Used for Case-based Reasoning

Attribute	**Description**
Threat	The description of the vulnerability where it includes a clause related to actual problem. Examples: “Code should not be dynamically injected and executed”, “Credentials should not be hard-coded”
Type	It can be a Bug, a Vulnerability or a Code Smell
Severity	Five severity levels are considered. **BLOCKER**: Bug with a high probability to impact the behavior of the application in production. The code must be immediately fixed. **CRITICAL**: A bug with a low probability to impact the behavior of the application. The code must be immediately reviewed. **MAJOR**: Quality flaw which can highly impact the developer productivity. **MINOR**: Quality flaw which can slightly impact the developer productivity. **INFO**: Neither a bug nor a quality flaw, just a finding.
Effort	The time (in minutes) estimate to fix the issue and update the tests.
Technical Debt	The time estimate to fix all maintainability issues (Code smells).
Language	The programming language that the issue occurs. For example : Javascript, Java, C#

Upon identifying and categorizing the security bugs, they are converted into *bug objects*. Each *bug object* contains the relevant details of bugs that will be the output of this component. The *bug objects* sent out by the Security Bug Pre-processor are transformed into Bug category objects. Ten Bug category objects are created with respect to OWASP T10 which contains details of each Bug object belongs to a particular category. Theses Bug category objects are sent to the Association Loader component.

C. Knowledge base and Association Inference Module

Association Loader used for querying the Knowledge Base. A Prolog converter is developed using SWI-Prolog to communicate with Java. Each bug category will be used to query the Knowledge Base, and the associated threat type

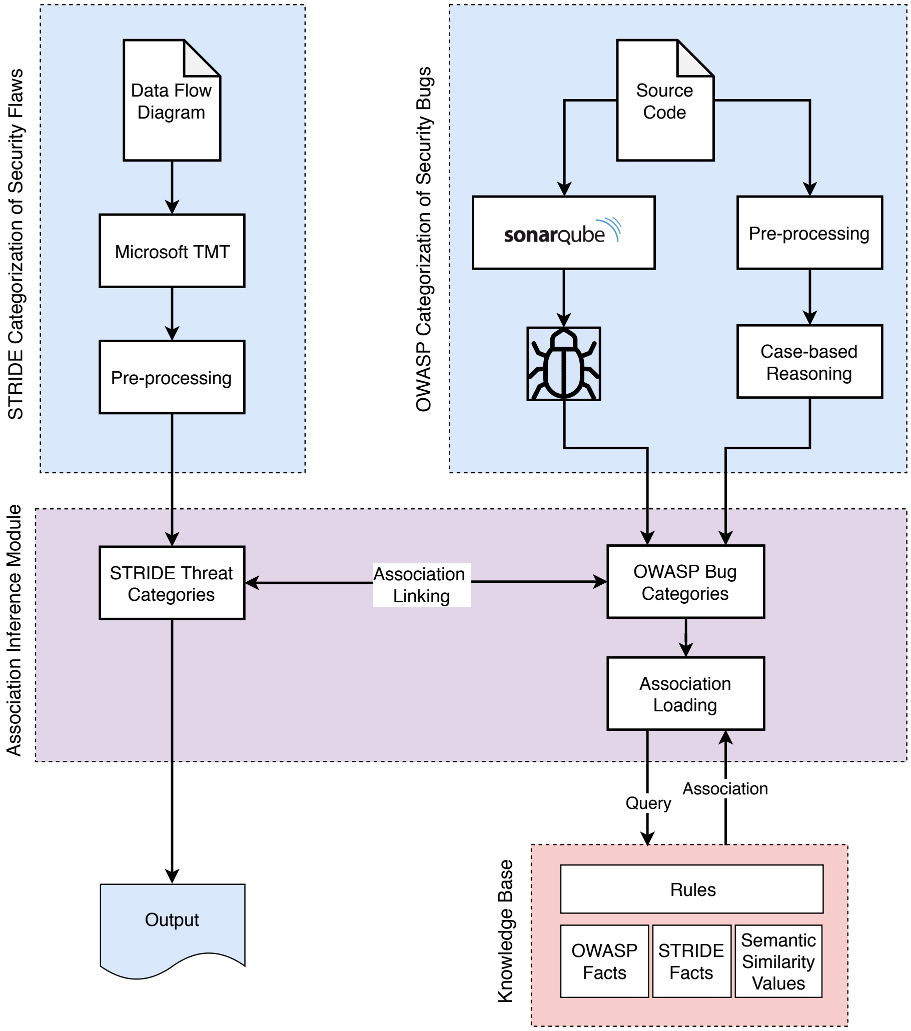


Fig. 3. Architecture of the Proof-of-Concept Implementation.

results held inside the Association Loader. The associated threat type results and the Bug Category Objects are sent to the Association Linker. Threat category objects from STRIDE Transformer and associated threat types and Bug objects from Association Loader will be the input to the association linker. After that, the Association objects generated. The Association objects are sent out from the Association Linker to the Output Builder.

The Knowledge Base is built using the SWI-Prolog. All the facts and rules described previously are contained in the Knowledge Base. The Knowledge Base has the capability of updating when the OWASP categories or Proactive controls revised. On the other hand, knowledge base explicitly allows expanding the knowledge contained in it using the additional knowledge of security experts.

V. Evaluation

The main focus of the evaluation is to find whether the potential root causes of an identified security bug lie in the design phase of the software application. Two case studies have employed for the evaluation process.

a) Case study 1 - User Authentication component: : Fig. 4 presents the DFD of the user authentication in a web-based application. It consists of two processes, one external entity, and a single data store together with associated data flows. In the evaluation process, threat modeling is conducted to identify the architectural-level security flaws, and static analysis is used to capture the security bugs at the implementation level. The association derived between security bugs and the threats are based on the security bug categories and threat categories. Table III depicts the possible threats identified by the threat modeling process. Similarly, through static code analysis, A2, A5, and A6 categories[3] of OWASP have captured. The results produced from the threat modeling process and the static code analysis provided as input to find the association between them. The derived associations present in Table IV.

[3] A1 to A10 are the top 10 threat categories of OWASP

Then, the highly relevant causes of the security bug categories were identified, and the corresponding countermeasures were applied to remove the security bugs in the source code. After repeating static analysis, it was observed that previously designated A2, A5 and A6 bug categories were removed successfully. Hence, it was evident that by removing the potential security specific root causes at the design level leads to resolve the security bugs at the code level.

TABLE III. IDENTIFIED THREATS OF THE USER AUTHENTICATION COMPONENT

Threat Type	No: of Threats
S	6
T	4
R	4
I	2
D	9
E	8

TABLE IV. ASSOCIATIONS DERIVED FOR CASE STUDY 1

OWASP Type	Derived association
A2	S, T, R, I, E
A5	T, R, I, E
A6	S, T, R

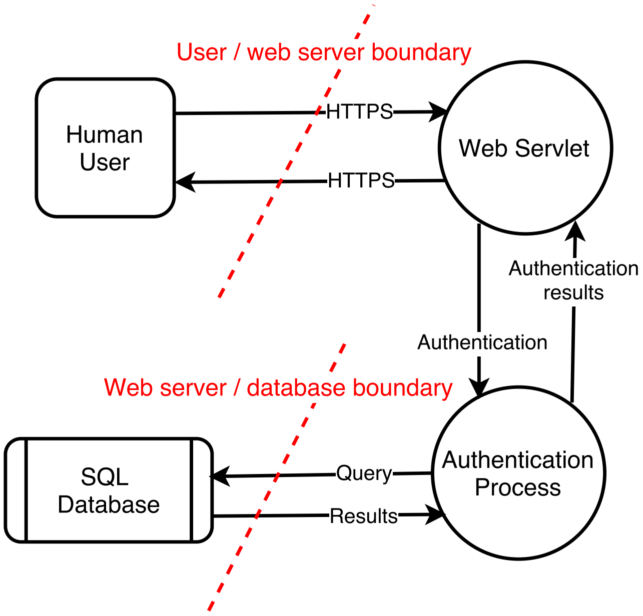


Fig. 4. DFD of the user Authentication Component of a Web Application.

b) Case study 2 - Large-scale web based application: : A large-scale industry project is selected to evaluate the scalability of the proposed approach. Fig. 5 presents a part of the DFD of the application. The full diagram is not presented due to its complexity with higher number of processes and data stores. Similar to Case study 1, DFD diagram is subjected to threat modeling and code base subjected to static analysis using the tools mentioned above. Summary of the identified threats presents in Table V. Based on the static code analysis, 26 security bugs related to A2 and A6 categories of OWASP have identified. The results produced from the threat modeling process and the static code analysis provided as input to find the association between them. The associations between threats and bugs that are derived from this approach is presented in Table VI. Then the corresponding countermeasures were applied to remove the security bugs in the source code. After repeating static analysis, it was observed that previously designated A2, A6 bug categories were removed successfully.

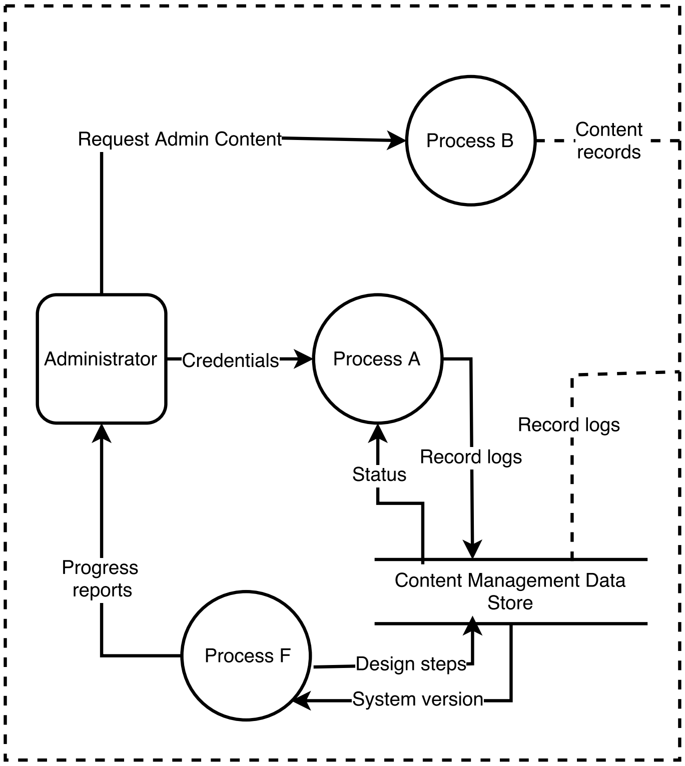


Fig. 5. Part of the DFD of the Large-Scale Web based Application.

TABLE V. IDENTIFIED THREATS OF THE LARGE-SCALE INDUSTRY APPLICATION

Threat Type	No: of Threats
S	12
T	0
R	0
I	5
D	5
E	5

A. Threats to Validity

The accuracy of the results obtained from the experiments depends on the analysis outputs given by SonarQube and MS Threat Modeling Tool. Despite the fact that the associations derived from this approach depict the possible causes for a security bug, even with such an association, pinpointing the

TABLE VI. ASSOCIATIONS DERIVED FOR CASE STUDY 2

OWASP Type	Derived association
A2	S, T, R, I, E
A6	S, T, R

exact location of the source code is improbable with a Level 0 DFD. Therefore, it is essential to consider the lower level DFDs at the threat modeling phase for efficient capturing of security bugs.

On the other hand, static code analysis tools may not be capable of capturing all the bug categories in OWASP Top 10. Hence, this approach is unable to derive associations for each OWASP Top 10 vulnerabilities contained in the source code. Therefore a manual code review is required to identify the remaining vulnerabilities. However, manual code reviews are not feasible with large-scale, complex projects.

VI. CONCLUSIONS

This paper presents a Knowledge-modeling approach to infer the associations among design artifacts and source code to reveal whether the root causes for security bugs lie in the design phase. This research employed a frame-based approach for the knowledge representation of security-specific information extracted from design documents and source code. Evaluation results imply that the knowledge-modeling approach successfully detects whether design flaws are propagated to the implementation phase. Besides, this paper provides experimental evidence of the usefulness and applicability of the concept of *Building Security In*. Moreover, this research contributes to the body of knowledge in secure software engineering by filling the research gap in interlinking security artifacts.

This approach has several limitations. First, security vulnerabilities at the design phase are detected solely based on DFDs, which is mainly due to the unavailability of tools to discover vulnerabilities of the other types of design artifacts such as UML diagrams. Secondly, the proof-of-concept implementation is entirely depending on the results produced by the Threat Modeling Tools and Static Analysis Tools. Thus, the derived associations are also could be biased to those tools.

The framework could serve as a stepping stone to the researchers in the field of software security, which was lacking previously. On the other hand, the knowledge base has provisions to evolve with different security aspects. As future work, the experiments are expected to repeat for large-scale open-source software systems. Furthermore, it is planned to improve this research to directly interlink security bugs with security flows by utilizing attack trees or case-based reasoning.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support provided by the National Research Council of Sri Lanka (Grant no: NRC 15-74).

REFERENCES

[1] S. Lipner, "The trustworthy computing security development lifecycle," in *Computer Security Applications Conference, 2004. 20th Annual*. IEEE, 2004, pp. 2–13.

[2] G. McGraw, *Software security: building security in*. Addison-Wesley Professional, 2006, vol. 1.

[3] M. Kreitz, "Security by design in software engineering," *ACM SIGSOFT Software Engineering Notes*, vol. 44, no. 3, pp. 23–23, 2019.

[4] G. Antoniol, G. Canfora, G. Casazza, and A. De Lucia, "Information retrieval models for recovering traceability links between code and documentation," in *Software Maintenance, 2000. Proceedings. International Conference on*. IEEE, 2000, pp. 40–49.

[5] A. Abeyratne, C. Samarage, B. Dahanayake, C. Wijesiriwardana, and P. Wimalaratne, "A security specific knowledge modelling approach for secure software engineering," *Journal of the National Science Foundation of Sri Lanka*, vol. 48, no. 1, 2020.

[6] A. Shostack, *Threat modeling: Designing for security*. John Wiley & Sons, 2014.

[7] R. Khan, K. McLaughlin, D. Laverty, and S. Sezer, "Stride-based threat modeling for cyber-physical systems," in *Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), 2017 IEEE PES*. IEEE, 2017, pp. 1–6.

[8] D. Wichers and J. Williams, "Owasp top-10 2017," *OWASP Foundation*, 2017.

[9] S. M. Srinivasan and R. S. Sangwan, "Web app security: A comparison and categorization of testing frameworks," *IEEE Software*, vol. 34, no. 1, pp. 99–102, 2017.

[10] M. Willberg, "Web application security testing with owasp top 10 framework," 2019.

[11] W. E. Zhang and Q. Z. Sheng, *Managing Data From Knowledge Bases: Querying and Extraction*. Springer, 2018.

[12] Y. Jia, Y. Qi, H. Shang, R. Jiang, and A. Li, "A practical approach to constructing a knowledge graph for cybersecurity," *Engineering*, vol. 4, no. 1, pp. 53–60, 2018.

[13] M. Howard and S. Lipner, *The security development lifecycle*. Microsoft Press Redmond, 2006, vol. 8.

[14] M. Frydman, G. Ruiz, E. Heymann, E. César, and B. P. Miller, "Automating risk analysis of software design models," *The Scientific World Journal*, vol. 2014, 2014.

[15] X. Yuan, E. B. Nuakoh, J. S. Beal, and H. Yu, "Retrieving relevant capec attack patterns for secure software development," in *Proceedings of the 9th Annual Cyber and Information Security Research Conference*. ACM, 2014, pp. 33–36.

[16] B. J. Berger, K. Sohr, and R. Koschke, "Automatically extracting threats from extended data flow diagrams," in *International Symposium on Engineering Secure Software and Systems*. Springer, 2016, pp. 56–71.

[17] L. Lambert, "Building security into your software development," Tech. Rep., 2018.

[18] S. S. Alqahtani, E. E. Eghan, and J. Rilling, "Sv-af—a security vulnerability analysis framework," in *Software Reliability Engineering (ISSRE), 2016 IEEE 27th International Symposium on*. IEEE, 2016, pp. 219–229.

[19] C. Wijesiriwardana and P. Wimalaratne, "Fostering real-time software analysis by leveraging heterogeneous and autonomous software repositories," *IEICE TRANSACTIONS on Information and Systems*, vol. 101, no. 11, pp. 2730–2743, 2018.

[20] M. Rath, M. Goman, and P. Mäder, "State of the art of traceability in open-source projects," 2017.

[21] C. Wijesiriwardana and P. Wimalaratne, "Software engineering data analytics: A framework based on a multi-layered abstraction mechanism," *IEICE Transactions on Information and Systems*, vol. 102, no. 3, pp. 637–639, 2019.

[22] E. Crifasi, S. Pike, Z. Stuedemann, S. M. Alnaeli, and Z. Altahat, "Cloud-based source code security and vulnerabilities analysis tool for c/c++ software systems," in *2018 IEEE International Conference on Electro/Information Technology (EIT)*. IEEE, 2018, pp. 0651–0654.

[23] R. Mahmood and Q. H. Mahmoud, “Evaluation of static analysis tools for finding vulnerabilities in java and c/c++ source code,” *arXiv preprint arXiv:1805.09040*, 2018.

[24] M. Abi-Antoun, D. Wang, and P. Torr, “Checking threat modeling data flow diagrams for implementation conformance and security,” in *Proceedings of the twenty-second IEEE/ACM international conference on Automated software engineering*. ACM, 2007, pp. 393–396.

[25] T. UcedaVelez and M. M. Morana, *Risk Centric Threat Modeling: Process for Attack Simulation and Threat Analysis*. John Wiley & Sons, 2015.

[26] D. Merritt, *Building expert systems in Prolog*. Springer Science & Business Media, 2012.

[27] A. Aamodt and E. Plaza, “Case-based reasoning: Foundational issues, methodological variations, and system approaches,” *AI communications*, vol. 7, no. 1, pp. 39–59, 1994.

SDN based Intrusion Detection and Prevention Systems using Manufacturer Usage Description: A Survey

Noman Mazhar[1]
Faculty of Computer Science and Information Technology
University of Malaya, Lembah Pantai
50603 Kuala Lumpur Malaysia

Rosli Salleh[2]
Faculty of Computer Science and Information Technology
University of Malaya, Lembah Pantai
50603 Kuala Lumpur Malaysia

Mohammad Asif Hossain[3]
Faculty of Computer Science and Information Technology
University of Malaya, Lembah Pantai
50603 Kuala Lumpur Malaysia

Muhammad Zeeshan[4]
School of Electrical Engineering and Computer Science
National University of Sciences and Technology
Islamabad

***Abstract*—Internet of things (IoT) is an emerging paradigm that integrates several technologies. IoT network constitutes of many interconnected devices that include various sensors, actuators, services and other communicable objects. The increasing demand for IoT and its services have created several security vulnerabilities. Conventional security approaches like intrusion detection systems are not up to the expectation to fulfil the security challenges of IoT networks, due to the conventional technologies used in them. This article presents a survey of intrusion detection and prevention system (IDPS), using state of art technologies, in the context of IoT security. IDPS constitutes of two parts: intrusion detection system and intrusion prevention system. An intrusion detection system (IDS) is used to detect and analyze both inbound and outbound network traffic for malicious activities. An intrusion prevention system (IPS) can be aligned with IDS by proactively inspecting a system's incoming traffic to mitigate harmful requests. The alignment of IDS and IPS is known as intrusion detection and prevention systems (IDPS). The amalgamation of new technologies, like software-defined network (SDN), machine learning (ML), and manufacturer usage description (MUD), in IDPS is putting the security on the next level. In this study IDPS and its performance benefits are analyzed in the context of IoT security. This survey describes all these prominent technologies in detail and their integrated applications to complement IDPS in the IoT network. Future research directions and challenges of IoT security have been elaborated in the end.**

Keywords*—*Intrusion Detection and Prevention Systems (IDPS); Internet of Things (IoT); Software Defined Network (SDN); Machine Learning (ML); Deep learning (DL); Manufacturer Usage Description (MUD)

I. INTRODUCTION

Internet revolutionizes our daily life and provides so many services that have become no more luxurious but the ultimate need for life. The Internet of things is the name of a smart environment that consists of many interconnected objects, to provide useful and instant services. These objects are not only traditional mobile devices or computers but also the gadgets of daily life like wearable devices, watches, and other smart articles. With the evolution of wireless sensor networks and recent improvement in the technology along with the expansion of low power devices, amplify the number of devices that can be connected to the Internet [1].

IoT inherits most of the conventional technologies for communication and so the security issues in them. The attacks such as worm attacks, denial of service attacks, etc. have become serious concerns [2]. There are many ways to address these security threats. A multitude of approaches used includes security systems and frameworks that have been adopted by the industry and research community. The intrusion detection and prevention systems (IDPS) is one of such systems.

The IDPS is not only capable to detect malicious activities like worms, viruses, distributed denial of services (DDoS), and others, but also capable to prevent the attacks before it happens. The detection system checks the traffic if it is normal or it should be blocked or regulate it to some different zone like a honeypot. Conventional IDPS has some limitations to defend against the latest security threats and also not feasible for devices with limited resources like IoT devices. Therefore, painstaking research has been done on developing the new generation of IDPS systems based on emerging technologies like Software Defined Networking (SDN) and Machine Learning (ML). A new contender, Manufacturer usage description (MUD), is also playing its role to reduce the attack surface for IoT devices.

SDN is recently a developing technology with different management and design approaches for networking. The design paradigm of this technology decouples the data and control planes. This gave the centralized and global view of the network. The controller is the decision-making authority while the switches and routers are the forwarding devices that handle data forwarding only. The controller and the forwarding devices work in a master and slave mode. The controller instructs the switches, how to handle the incoming and outgoing packets or flows. SDN is considered to be the best network model to address the heterogeneous changes in the overall network [3].

Along with the SDN technology, a new concept has been introduced for the identification of IoT devices known as "Manufacturer usage description" (MUD). MUD is a developing concept to define IoT device behaviour for network communication [4]. This automatically identifies the device and helps the security system to figure out the abnormal or malicious nodes within the network. For complete detection and monitoring of malicious activity in the network machine learning plays its role. For the detection of malware and malicious traffic, ML techniques have the primary role. In traditional networks, detection of malicious traffic and classification of a network attack is achieved using predefined rules and specifications which are limited to address new kinds of attacks. The main application of using ML in SDN networks is the control of the entire network rather than just focusing on localized policy or certain rules [5]. Such techniques show great potential for network traffic classification and solving prediction problems [6]. ML is used in the IDPS systems for the detection of security attacks and to predict future threats to the system.

In this paper, our research focuses on SDN based IDPS systems for IoT security using ML and device profile based techniques like MUD. Further, the study classifies IDPS systems based on the technology they use. Besides, we compare the conventional IDPS systems to new generation IDPS systems in the context of IoT security. Also, the study shows the performance of these new generation IDPS for IoT based networks. The overall research approach is shown in Fig. 1, we provide the future directions for upcoming secure systems along with the IDPS future perspectives in the domain of IoT security framework.

A. Contribution of this Survey Article

As the IoT scalability and heterogeneous increases over time, IDPS systems become inevitable. There is a lot of survey work done on the intrusion detection systems for the IoT but to the best of our knowledge, no work has been done on IDPS for IoT devices using the device profile base techniques like MUD for comprehensive security for IoT. As shown in Fig. 3. The contribution of this paper is to analyze the end to end IoT security solution based on IDPS using techniques like SDN and ML. The main points are given below:

1) Present the taxonomy of the IDPS for IoT using SDN and ML and hybrid approaches.
2) Investigate the IoT device profile standard like MUD in enhancing the IoT security in IDPS for IoT.
3) We analyse the performance of the IDPS systems based on SDN, ML, and MUD.

The paper is organized as shown in Fig. 2. Section II provides an overview of IoT, several security issues of IoT and their taxonomies, IDS and IPS and their integration. Section III presents the basic overview of SDN, ML and MUD technologies. Section IV describes the detail applications of SDN, ML and MUD in IDPS of IoT system. Section V outlines some open issues, challenges, and future research directions, and finally, Section VI concludes the paper. The acronyms used in this paper and their full forms are listed in Table I.

B. Related Works

Machine Learning applications have been proliferated in almost every field, especially in security. This gets the attention of many researchers and industrialists'. But ML requires a platform to unleash its potential. For network security IDPS provides a strong platform. Powering the IDPS system with ML, SDN, and MUD proving to be useful. IDPS system becomes a lot more effective and worthy for network security as in paper [7] presented an updated review on the IDPS systems. It shows the classification of IDPS systems and their role in securing the conventional network. The review did not talk about IoT security using IDPS systems. It also did not discuss SDN. Going further, paper [8] emphasized on the intrusion detection system in general and then specific to the context of IoT. Since IoT devices and systems are so diverse therefore they require proper security mechanisms to defend the system from cyberattacks. The study works more on the IDS for IoT networks and does not put any light on the latest technologies prevention techniques.

Much research focuses on the IoT security aspects like in [9] the author mostly focused on the IoT protocols and standards for different layers of network stack like medium access layer, Network Layer, and Session layer. Other than this the study explains different management and security standards, developed by the international engineering task force (IETF), Institute of electrical and electronics Engineers (IEEE), international telecommunication union (ITU), and other bodies. But the survey is more focused on IoT security, but not much explains about the new technologies like SDN and ML and their role in IoT security. Similarly, a study [10] focused on the state of art approaches like ML for IoT as well as intrusion detection for network security. But the focus of the study is on the ML approach and did not cover other technologies like SDN and MUD in this case. The study [11] explained the IoT architecture, IoT attacks. Then explain IDS technology and its types describing its use in IoT networks. It also classified different IDS systems used in IoT networks and the type of technology they use. Again, the issue is that they didn't consider the latest techniques like SDN and MUD for the security of IoT networks. Furthermore, they shed no light on the prevention systems along with the detection systems.

Further, the research shows the applicability of intrusion detection systems in IoT security as in research [12] showed the importance of IDS in defending IoT devices from cyber threats. It classifies different IDS systems based on the detection and deployment scenarios, explains different IoT attacks, and compares different IDS systems against the detection accuracy, false positive, resource consumption, and other attributes. The research work considers different IDS for the IoT defence systems but they were mostly based on conventional techniques, no state of art technique was considered like SDN and MUD techniques.

Next, the SDN comes into play and combining other techniques like ML provide security solutions as shown in a study [13] surveyed ML/DL techniques used in the SDN based IDS system. Also, the paper evaluates different deep learning techniques to analyses their impact on network security. The study evaluates that with the ML/DL approach there is a problem of the dataset for more accurate results. Also, with SDN the centralized controller is a bottleneck when we need to

TABLE I. SUMMARY OF MAJOR ACRONYMS USED

Acronym	Description	Acronym	Description
ACL	Access control list	MUD	Manufacturer Usage Description
CIAA	Confidentiality, Integrity, Authenticity and Availability	NFB	network function virtualization
CNN	convolutional neural network	RFID	radio frequency-based identification
DDoS	distributed denial of service	RNN	recurrent neural network
DL	Deep learning	SDN	Software Defined Network
IDPS	Intrusion Detection and Prevention Systems	SIEM	security incident and event management system
IDS	Intrusion Detection Systems	SOM	self-organizing maps
IoT	Internet of Things	SVM	support vector machine
IPS	Intrusion Prevention Systems	URL	Uniform Resource Locator
ML	Machine Learning		

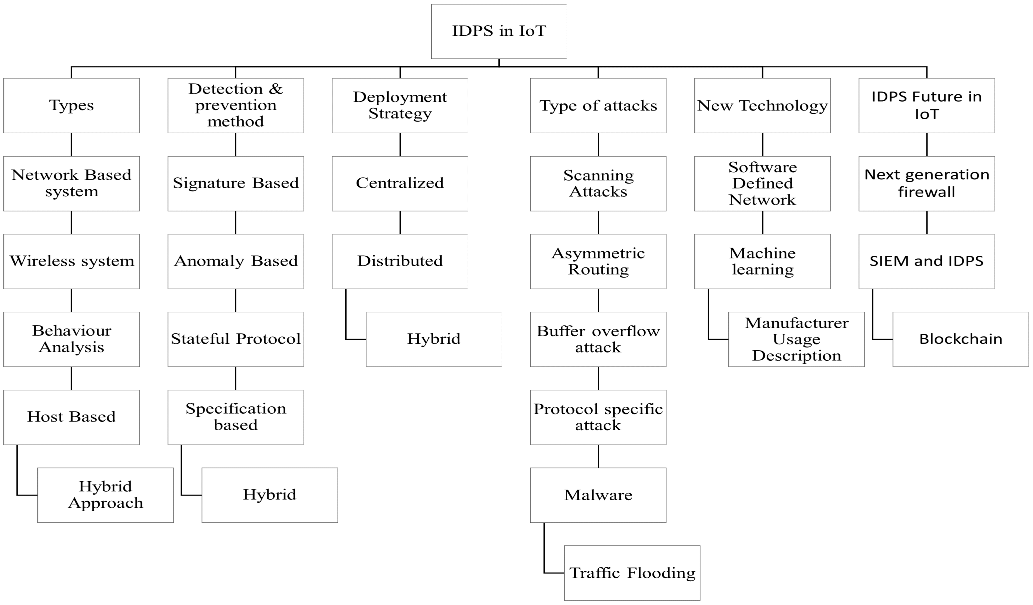


Fig. 1. Taxonomy of IDPS in IoT

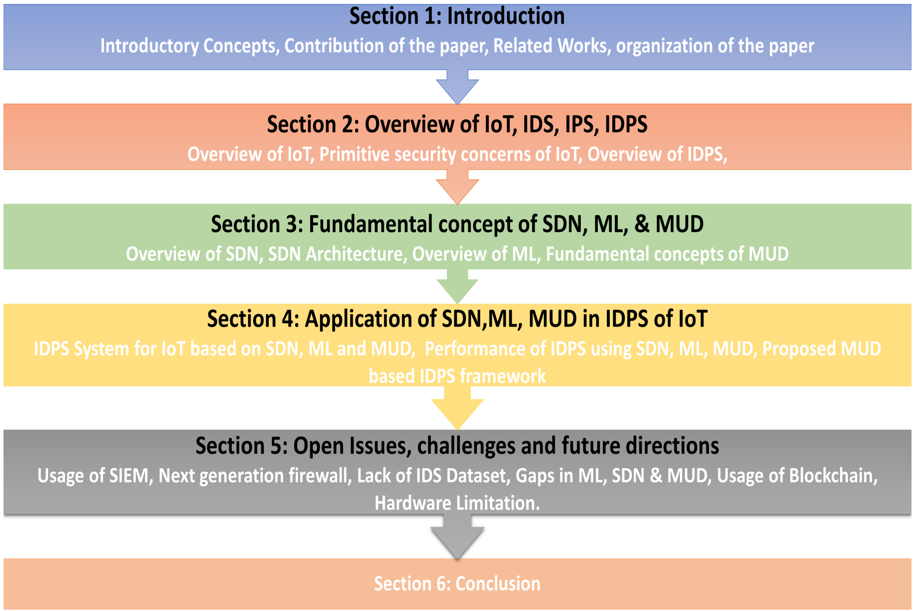

Section 1: Introduction
Introductory Concepts, Contribution of the paper, Related Works, organization of the paper
Section 2: Overview of IoT, IDS, IPS, IDPS
Overview of IoT, Primitive security concerns of IoT, Overview of IDPS,
Section 3: Fundamental concept of SDN, ML, & MUD
Overview of SDN, SDN Architecture, Overview of ML, Fundamental concepts of MUD
Section 4: Application of SDN,ML, MUD in IDPS of IoT
IDPS System for IoT based on SDN, ML and MUD, Performance of IDPS using SDN, ML, MUD, Proposed MUD based IDPS framework
Section 5: Open Issues, challenges and future directions
Usage of SIEM, Next generation firewall, Lack of IDS Dataset, Gaps in ML, SDN & MUD, Usage of Blockchain, Hardware Limitation.
Section 6: Conclusion

Fig. 2. Organization of the Paper

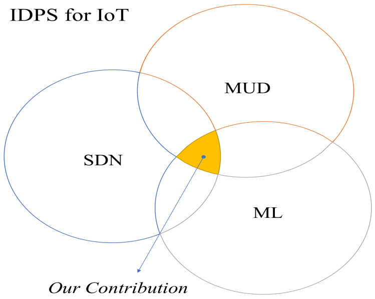


Fig. 3. Contribution of this Paper

do real-time intrusion detection. The study, however, focusses on ML and SDN based IDS but did not give any statistical analysis of their performance. Also, it did not explain the effectiveness of these IDS systems in the context of low power and resource-constrained devices like IoT devices. In this [14] the author focused on the SDN and its application to secure the computer network. The study explains SDN issues like scalability, resilience, security, incorporation with the conventional network. Further, the author shows a little description of the IDS based on the SDN and what type of IDS is more effective in securing the network against the threats. The research [15] showed the role of SDN in IoT to defend against the DDoS attacks. Discuss the different detection techniques that are possible in the SDN for attack detection like ML, traffic analysis, and connection-oriented. However, the study only uses ML techniques for the detection of attacks and did not consider other techniques like signature-based, specification-based, and stateful protocol analysis. It also did not describe any authenticated model for IoT devices like MUD.

However, a new dimension for security has been discussed in the research [16] suggested that the traditional security techniques where security is provided as a pre-emptive measure against known attacks are not sufficient for future attacks on IoT devices. The study proposes a secure by design thinking, for proactive defence system rather than passive systems. This paper has not given much description of the detection system as they are gaining importance in IoT security due to the heterogeneity and scalability of IoT devices. It also did not cover much about the authentication technique in IoT security. The summary of the related works and additional contributions of our paper compared to those related works is given in Table II.

II. Overview of IOT, IDS, IPS, IDPS

This section describes the basic overview of IoT and its various security concerns. It also discusses the concept of IDS, IPS, and IDPS along with their limitations.

A. *Overview of IOT*

IoT objects are intelligent devices, not dump objects. These smart things use different communication mediums for interacting with each other and outside world over the Internet. These intelligent things have certain properties in common as shown in Fig. 4 [17].

Identification: is required for every device to communicate with each other, IPv6 protocol can be used for this purpose.

Sensing: is the capability of the device to get some physical world data.

Communication: is the ability of the device to be able to communicate with the user and the other devices in the network and outside world.

Computation: is required for information processing.

Services: is the functionality provided to the users by these devices based on the data they acquire from the outer world.

Semantics: is the concept that the devices are supposed to get the right information from the environment and give them services in a timely fashion. Example for these devices are beagle boards [18], [19], Arduino [18]–[20], cubie Board [20], [21], Raspberry Pi [20]–[22] and radio frequency-based identification (RFID) [23]. There are some security concerns as discussed in the coming sections.

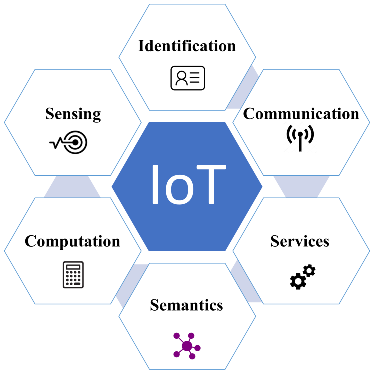


Fig. 4. IoT Device Attributes

1) Primitive Security Concerns of IOT: IoT security has become a big challenge for the industry and academia. With every passing day, new devices come in the market with a plethora of new useful applications. But this exposes more risk towards the security and privacy of the data. Before going to discuss IoT security threats first few basic security requirements are described below known as CIAA (Confidentiality, Integrity, Authenticity, and Availability) [8], [15], [24].

Confidentiality: is the concept that assures no unauthorize service should access the private information and it maintains the privacy and proprietary of the information.

Integrity: is the concept in which the information of the IoT devices should not be modified by any unauthorized user or object.

Authenticity: is the concept that validates the fact that the partner involves in the information transaction is genuine and as in the same what they claim to be.

Availability: is the feature that determines the service is available to the user when and where it is required. In this context, all the storage, processing, and communication medium should work reliably.

TABLE II. Related works and comparisons with our paper

Reference	IDS	IPS	ML	SDN	Signature	Anomaly	Specification	Entropy	Hybrid	topics not covered
Azeez, et al. [7]	✓	✓	✓	✗	✗	✗	✗	✗	✗	IDP
da Costa, et al. [10]	✓	✗	✓	✗	✗	✗	✗	✗	✗	SDN, MUD
Tiwari and Mishra [8]	✓	✗	✗	✗	✓	✓	✗	✗	✗	SDN and authorization techniques.
Sultana, et al. [13]	✓	✗	✓	✓	✗	✗	✗	✗	✗	IDP and IDS for IoT networks
Sahay, et al. [14]	✓	✗	✓	✓	✗	✗	✗	✓	✗	SDN based detection for IoT
Pajila and Julie [15]	✗	✗	✓	✓	✗	✗	✗	✗	✗	Focus on IDS systems
Hajiheidari, et al. [12]	✓	✗	✓	✓	✓	✓	✓	✗	✗	SDN based IDS and authorization techniques
Choudhary and Kesswani [11]	✓	✗	✓	✗	✓	✓	✓	✗	✗	Classification of IDS systems for IoT
Restuccia, et al. [16]	✓	✗	✓	✓	✗	✗	✗	✗	✗	IDPS for IoT networks Using MUD

2) IOT Attack Vectors: The Open Web Application Security Project (OWASP) has published a detailed draft regarding the attack surfaces of IoT, these are the areas in IoT systems and applications that are vulnerable and prone to threats. Fig. 5 shows a summary of these attack surface areas [25].

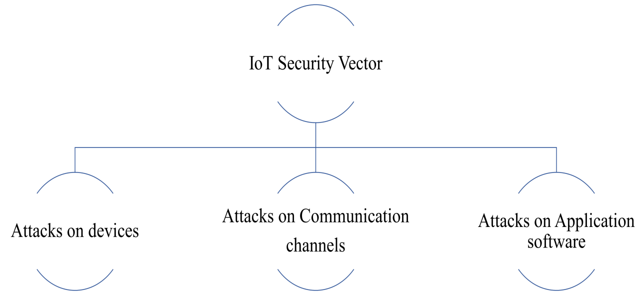

IoT Security Vector
Attacks on devices
Attacks on Communication channels
Attacks on Application software

Fig. 5. OWASP Taxonomy of IoT Security

Attacks on Devices: devices are the primary sources from where the attack can be initiated. The main parts of the device like memory, firmware, physical interface, web interface, and network surfaces are vulnerable. The attackers can further exploit the default setting, old components, and insecure update mechanism.

Attacks on Communication: communication channels are another area of security concern. The channel connects the IoT devices and the outside world. The protocol used for communication in the IoT networks has security issues and can affect the entire system. IoT systems get vulnerable to attacks like Denial of the Service (DoS) and spoofing.

Attacks on Application software: in this the vulnerabilities of web application and software used in IoT systems can become a great cause of a compromised system. The web application can steal user data and insert malicious updates in the system. To defend these challenges, many techniques have been developed and are being used in the industry, intrusion detection and prevention system is one of them.

B. Overview of Intrusion Detection and Prevention System

The working principle of the IDS is to monitor the data packets to determine abnormal traffic. There are three major parts of IDS system monitoring, analysis, and detection [26]. The core modules of IDS are the analysis and detection based on the algorithm, which also generates the alarms on the detection of any intrusion [27]. As the detection systems become more common nowadays, attackers find covert ways to exploit the loopholes in the system, like bypassing the system and disabling the system. This results in a DoS kind of attack. To mitigate these kinds of attacks, the researchers suggest the IDPS. This system is not visible to the attackers thus restricts the communication to the other components of the network [7]. The taxonomy of IDPS in IoT is given in Fig. 1.

The IDPS system can be classified using different perspectives as shown in the figure. One way is the application-specific types of IDPS like for network-based applications, wireless applications, behaviour analysis based applications, Host-based and hybrid approaches. Next, we can classify the IDPS based on the detection and prevention techniques used by them like anomaly, signature, specifications, protocol analysis, and hybrid approaches. One of the important aspects of such systems is the deployment strategy, like centralized and distributed deployment of IDPS systems. If required hybrid mode deployment can be used according to the application requirement. One interesting IDPS classification is regarding the security coverage provided by the IDPS systems.

Complete range of security coverage is not discussed here, however, three main areas are discussed like security coverage for IoT devices, communication medium, and application layer. Further, the IDPS systems are discussed in the context of the conventional use of technologies versus new state of art technologies. IDPS constitutes of many building blocks like the type of data network used for IDPS, protocols for the communication, detection and prevention techniques, and application software.

The general IDPS architecture contains sensors, firewalls, management server and console as shown in Fig. 6. Typical

IDPS systems use sensors in the network to monitor the network traffic. At the input of the network, there is a firewall installed as the first line of defence after that there is IDPS sensor which monitors incoming network traffic and passes the information to the management server and the console, while at the same time, it sends network traffic to the local network. Many security techniques discussed above put a focus on the defensive strategy against the attacks.

However, there is a need for the method to detect the ongoing attack. Heavy and complex antivirus and firewalls cannot be used in IoT devices. In this context, lightweight IDS has been devised [28]. One of the examples of attack detection is to monitor different parameters like CPU usage, storage usage, and throughput usage, etc. [24]. Another way to check the energy profiles of power consumption may lead to detecting attacks [29]. Many anomaly detection techniques can be used like the one described in [30] to test the packet drops, send rate, and signal strength. One of the latest methods is to use an ML approach for intrusion detection. Like in [31] random forest technique has been used for the monitoring of traffic flows. An attack is detected when some flows exhibit not according to the normal standard. The goal of IDS is to detect any intrusion in the network or detect any malicious node. Also, it alerts the users about it timely. IDS works like an alarm system and monitors the whole system and generate an alarm when some malicious activity is observed. It gave protection from internal and external threats. The following sections discuss different aspects of IDPS in detail.

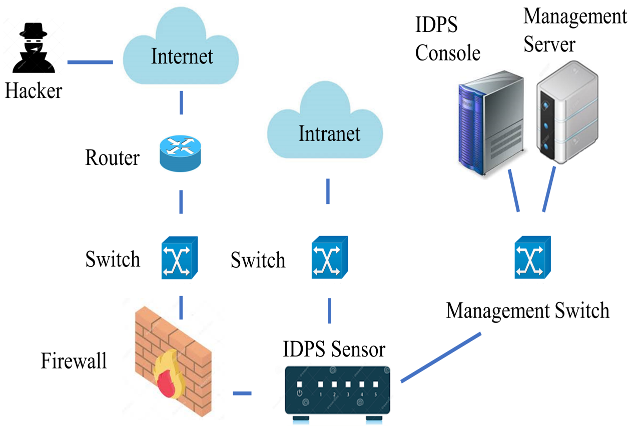


Fig. 6. Working Principle of IDPS

1) IDPS based on application type: The IDPS can be classified based on the kind of applications in which they are used [32]. They are [33] discussed as follow:

Network-Based IDPS (NIDPS) this type checks the network traffic for devices or specific network and application protocol for the detection of malicious traffic. It can detect many interesting events in the network. The deployment of these systems is mostly at the edge of networks like the boundary of router or firewall, remote access servers, virtual private network servers, and wireless networks.

Wireless IDPS (WIDPS) these devices monitor the wireless network traffic and the wireless network protocols for monitoring of suspicious activity. The WIDPS compose of the almost same components like the network IDPS like database servers, sensors, and management servers. The only difference is sensors in both types of IDPS. The wireless sensors have to perform more complex functions of wireless networks.

Network Behavior Analysis (NBA) this system analyzes the network traffic and its statistics to determine malicious traffic like DDoS, malware, and any policy violation in the network. The NBA components consist of consoles, sensors, and analyzers.

Host-Based IDPS (HIDPS) this type of system is designed to monitor a single host and all the activities within the single host. These kinds of systems monitor both wired and wireless connections. Most of the host-based systems use agents that are installed on a single host for the detection of malicious activity.

2) Detection Type: Based on the detection method, there are normally four types of IDS. Each one of them is explained below.

Signature-Based: this type of system detects the attack by determining the incoming network behaviour and match it with the attack signature stored in the internal database. If the match occurs then an alert is generated. It is also known as rule-based detection. This compares the present profile of the network from the previously-stored profiles containing different attacks [34]. This type of IDPS is accurate in determining the known attacks whose signature is stored in the database. But for new kinds of attacks, this approach is not suitable because the matching signature of attack is not available [35], [36].

Anomaly-Based: this type of IDS defines normal network behaviour and any activity not conforming to this is an intrusion or threat [34]. If the network behaviour deviated from the normal profile it generates alert. This approach is good for the detection of new attacks. But this type of IDPS has high false-positive [32], [37], [38].

Stateful Protocol Analysis: this type of system works by comparing the established profile of the protocol and against their behaviour. This standard behaviour is provided by the vendor, just like the signature-based system that compares the behaviour from the given list. This stateful analysis of protocol requires a very deep understanding of the protocol and applications they interact with [39].

Specification-Based: this type of IDS is based on the physically defined polices by the users. Any activity that is against such rules is considered a threat to the system [34]. It is the system in which set of rules and their thresholds are defined for all the network elements like nodes, protocols, and routing tables. If the network behaviour deviates from the set specifications then the threat is detected.

Hybrid Approaches: this approach combines all these types of systems specification-based, signature-based and anomaly-based systems to get the maximum advantage from their strength and lesson the weakness in those systems. SVELTE is an example of a hybrid system [40]This system combines the signature-based and anomaly-based methods.

3) Deployment Strategy Type: We discuss different ways to deploy the IDS in a network, based on the security application scenario.

Distributed IDS: In this strategy, the IDS is placed in every node. IDS needs to be efficient and optimize for the IoT devices due to resource constraints.

Centralized IDS: In this kind of strategy, the IDS is placed at the centre of the network such as at the border router or some specific host. All the Internet traffic from the low power and lossy networks (LLN) nodes pass through the border router, therefore the border router can monitor all the traffic from the Internet and the LLN nodes [40], [41].

Hybrid Approach: In this approach, both methods are followed like a centralized and distributed approach to have the benefits of both approaches. One of the practical approaches for this is to transform all the network into regions and clusters so that each sub region and the cluster can have one node that acts as a host and responsible to monitor all the nodes in the cluster.

C. Conventional IDPS Systems for IoT

Since the IoT devices are resource constraints, therefore conventional IDS system is not suitable for them. A hybrid scheme is required to suit the needs of these resource-limited devices. One of the studies [42] concluded that in signature-based IDS the headers of the incoming packet are compared to the set of the rules. If the number of packets increases the computation will more CPU cycles. Therefore, signature-based IDS are not suitable for IoT devices. Dynamic encoding scheme [43] uses a distributed signature-based IDS system among the ubiquitous sensor networks based on IP, making it lightweight and suitable for low power devices. The study proposed an outlier algorithm TAOOD for the fact that the majority of IoT devices have power constraints and have low quality. So, the resource limitation must be taken into consideration of the outlier algorithm. The TAOOD "Tolerance based adaptive online outlier detection" technique shows higher performance in terms of accuracy to the tolerance and outlier parameters.

The Finite State Machine (FSM) [44] approach has been used to detect the rank and local repair attacks. However, research [45] shows that resource constraint and heterogeneity the IoT networks are causing vulnerability in IoT networks and become a cause of too many threats. So, the study proposes an algorithm based on anomaly detection at the perceptual layer. The algorithm uses anomaly mining techniques. Similarly, a system [46] uses the artificial immune technique in the IoT environment. The study uses the concept of immature, mature, and memory detectors for attack detection.

Service-based approaches [47] presents an architecture constitute of services known as service-oriented architecture SOA used in the IoT system. It uses to prevent the DDoS attacks using learning automata concept. However, conventional methods are not adequate for IoT security [48], so an artificial immune system concept has been used containing the antigen, non-self, and detector in an IoT environment. Hybrid scheme [49] uses a two-layer strategy for protection. One is cryptography and the other is the IDS technique. TESLA [50] protocol is used for DoS attack prevention, as it uses few resources.

Some network-based approaches [51] address the DoS detection using IDS architecture based on the IDS probe. The IDS monitor the sixLoWPAN traffic for the detection of attacks. However, the SVELTE [40] a real-time IDS implemented in Contiki OS. This IDS also works for 6LoWPAN based border router. Some system uses a "security incident and event management system" (SIEM) and "frequency agility manager" (FAM) [52] for the detection of flooding, jamming, and DoS attacks.

Performance of the detection system is very important, a new lightweight protocol known as heartbeat protocol [53] has been introduced in the IDS of IoT networks. Also for real-time detection, a proposed scheme [54] event-based IDS system for real-time detection in IoT networks has been introduced. The IDS works on the event processing model (EPM). The model proposes a rule base scheme, the rules are stored in a repository. Device authentication is another way to protect the devices, a research [55] presents an authentication scheme for IoT devices using XOR manipulation. It is used for counterfeiting and privacy protection of the IoT devices.

Attack categorization is also important, as this research [56] works on the categorization of the Sybil attack in the social aspect of the IoT. The study proposes defence schemes for Sybil using "social graph-based Sybil detection" (SGSD) and "behaviour classification-based Sybil detection" (BCSD). Similarly, [57] shows IDS capable of detecting wormhole attacks and the attacker uses the Contiki OS and Cooja simulator. The IDS can be used in a centralized and distributed scenario. Another research [58] purpose an intrusion detection system against the sinkhole attack known as "intrusion detection of sinkhole attacks on 6LoWPAN for the Internet of things" (INTI).

RFID can be used for authentication of IoT devices, as an interesting study [59] used the RFID authentication mechanism for IoT devices. Also, propose the elliptical curve cryptography (ECC) based on protocols using RFID authentication and the author propose three extended RFID protocols based on ECC. This study [60] proposes a cloud-based antimalware system known as CloudEyes capable of providing reliable and efficient security for resource constraint devices.

Some modelling techniques [61] describes a behavioural modelling IDS (BMIDS). This IDS based on immunity inspired algorithm to distinguish between behaviour changes. The use of artificial intelligence is also quite useful against cyber threats, as a work [62] proposes IDS based on the neural network approach to detect the DDoS and DoS. In the same direction, this work [63] uses compressed sensing technology to monitor the network data, also uses the "support vector machine" (SVM) for the detection of anomalous compressed data. The author [64] proposes a semi-auto building to protect the network topology based on RPL using a specification-based IDS system. The model can detect attacks on topology and RPL like a wormhole, black hole, and selective forwarding. Table III shows the summary of the conventional IDPS systems for the IoT systems as discussed above, it also shows the scheme and the techniques used in such systems along with the security impact of such systems.

TABLE III. CONVENTIONAL IDPS SYSTEMS

References	Signature	Anomaly	Specification	Hybrid	Other	Results	Remarks
Amin, et al. [42]	✗	✗	✗	✓	✗	Work well for a large number of the signature set.	Has limitation to detect new kind of attacks
Shen, et al. [43]	✗	✓	✗	✗	✗	TAOOD perform better than normal sliding window based detection.	The scheme is restricted to detect few attacks.
Le, et al. [44]	✗	✗	✓	✗	✗	Detect the attacks with some processing overhead	No real-time implementation has been done.
Fu, et al. [45]	✗	✓	✗	✗	✗	Theoretical results are 100% detection.	The system is not tested in real-time scenarios.
Liu, et al. [46]	✓	✗	✗	✗	✗	Shows better performance over the other conventional detection systems.	The study does not show what new types of attacks the system can detect.
Misra, et al. [47]	✗	✗	✗	✗	✓	Using LA shows better results to detect DDoS as compared to without LA.	LA requires a lot of processing power, also to detect the runtime attacks will be a question mark for the system.
Chen, et al. [48],Le, et al. [49]	✗	✗	✗	✓	✗	———	The system didn't show any actual results.
Kasinathan, et al. [51]	✓	✗	✗	✗	✗	The results show detection of flood attacks increase as the nodes increases.	The system is not implemented in an actual scenario.
Raza, et al. [40]	✗	✗	✗	✓	✗	True positive rate is 100%	Need to test against a broader range of attacks
Kasinathan, et al. [52],Jun and Chi [54]	✗	✗	✓	✗	✗	AS the traffic increase the system Akeeps normal processing as compare to traditional IDS.	Require more resources to process CEP.
Pongle and Chavan [57]	✗	✗	✗	✓	✗	The attack detection is more than 90%.	Require more attributes of the nodes other than location to detect more attacks.
Arrington, et al. [61],Chen, et al. [63]	✗	✗	✗	✗	✓	The results show efficient detection results.	Now in the future trend behavior attribute is bypassed using smart techniques.

III. FUNDAMENTAL CONCEPT OF SDN ML & MUD

A. Software Defined Network

Traditional computer networks consist of many devices like switches, routers, middleboxes, servers, and hosts. Network operators are responsible for the configuration and maintenance of the network. As the network configuration requires manual low-level device-specific syntax to configure the network, making it more complex, time-consuming, and error-prone. This becomes the big reason for the network downtime [65], [66]. To address these issues software-defined network (SDN) technology emerges. SDN decouples the control plane from the data plane of the network devices [67]. All the intelligence is centralized in SDN and the device is known as a controller, the rest of the devices like switches, routers, etc. are known as forwarding devices and come under the category of the data plane. The devices at the data plane use flow rules to handle the incoming packets. The rules are programmed in these forwarding devices from the controller using a standard interface such as OpenFlow [68]–[70]. In addition to this, the controller can get the flow information of all the devices attached to the network to provide the network administrator's global network status [71]–[74]. This makes network management simple and transparent.

1) SDN Architecture: SDN Architecture is composed of three layers and three interfaces [75] as shown in Fig. 7

Application Plane: This layer contains the business applications that use SDN communication and network services. Examples of such applications are security application, management, and monitoring applications and network virtualization applications.

Control Plane: This plane consists of SDN controllers responsible for the control of the network and defining the network forwarding behaviour using open interfaces. The controller uses three interfaces northbound, southbound, and east/westbound interfaces. The northbound interface helps the developers to develop the SDN application while hiding the lower layer details is known as the northbound interface. The southbound interface provides the communication between the SDN controller and the data devices like switches and routers at the data plane. It also specifies the communication protocol between the controller and the data plane devices. While the Westbound API's are responsible for the controller communicates with the legacy network devices.

Data plane: Consist of network devices like router, switches, IDS, and firewall their main responsibility is to forward the data and filter it.

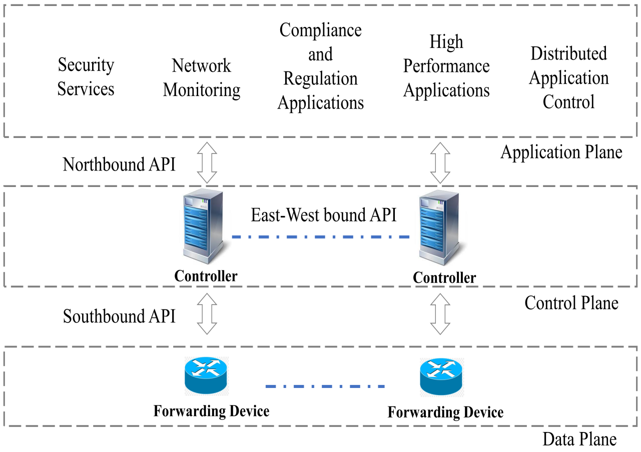


Fig. 7. SDN Architecture

B. Machine Learning

The concept of ML is to make machines to learn automatically from the given data [76]. Also, detect patterns without explicitly programmed the device [77]. The ML algorithm classification is based on the type of data they learn from and what function they perform [77]. Mainly the ML approaches can be categorized in supervised learning, unsupervised learning, and semi-supervised learning [78]. The ML classification is shown in Fig. 8.

Supervised learning: In this the algorithm input the label data, base on this it will predict the unknown cases. Random forest and support vector machine is the example of this type of algorithms used for regression and classification problems respectively [76]. Support vector machine (SVM) and random forest both are used for network intrusion detection system (NIDS). However, SVM is more resource-hungry in the context of memory and computational power [77].

Unsupervised Learning: This kind of algorithm gets the unlabeled data, they learn from the data distribution and data pattern to determine the unknown data [76]. Examples of this are the principal component analysis (PCA) and self-organizing maps (SOM). PCA is mostly used for feature extraction before applying the classification [79]. Other algorithms based on the clustering technique as K-means and other distance-based algorithms are used for anomaly detection. SOM is developed to reduce the payload in NIDS [80]. The main disadvantage of the clustering algorithm is its dependency on the initial conditions like centroid that produces high false-positive results [81].

Semi-Supervised Learning: In this learning process there is a small portion of labelled data and a large chunk of unlabeled data. It is a useful scheme when a large amount of data is unlabeled. For example, photo archives where few images are available [82]. A semi-supervised support vector machine is used for the improvement of the NIDS [83], [84].

Deep Learning algorithms an update on the artificial neural networks use the computation available [85]. The DL technique permits the algorithm to represent the data at various levels of generalization. The main application is object detection, detect the network intrusion detection and many different domains [86]. DL can be learned and trained both ways supervised and unsupervised [76]. Deep learning in a supervised way has an example of a convolutional neural network (CNN). The main application of the CNN is in face recognition and 2D images [86], [87]. DL in an unsupervised way has the example for autoencoder [88] used to learn the representations for the application of dimensionality reduction. A deep belief network (DBN) [89] is another example that is learned using unlabeled and labelled data for the feature extraction and classification purpose. DL in a supervised and unsupervised way has an example recurrent neural network (RNN) [90]. Speech recognition is the main application of the RNN algorithm.

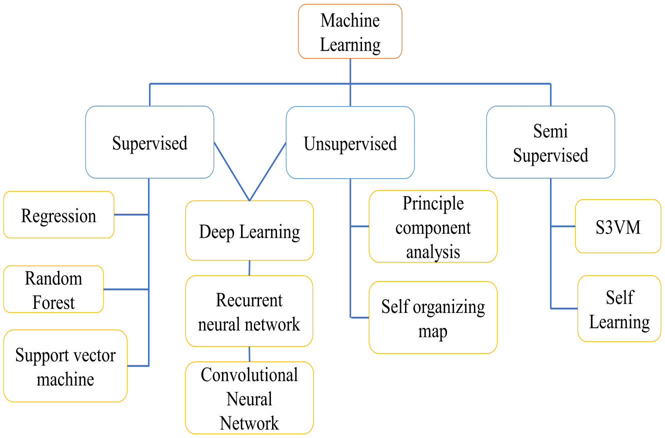


Fig. 8. Taxonomy of Machine Learning

C. Manufacturer Usage Description (MUD)

IoT devices are specialized devices with specific tasks to perform. These devices need to communicate over the network and have special communication requirements that the vendor only knows. For example, the printer only requires to print on the LPT port and the local access port for HTTP is 80. Therefore, it should deny all the other means of access to the network. For this purpose, an idea of MUD [91] has been introduced. MUD declares the intended communication pattern by the manufacturer for the network infrastructure using a network access control list. The MUD works as shown in Fig. 9.

The IoT devices also called things wanted to join the IoT network, they emit Uniform Resource Locator (URL) to the MUD manager. The manager sends this URL to the MUD server. The server sends the device profile file to the MUD manager. Based on the device profile, the manager configures the switch and installs the policy for the device. After this verification process, the device got permission to enter the network. MUD provides a defence system against the malicious agents that got into the network and try to launch an attack on the network infrastructure. This technology also provides defence against compromised devices to attack other devices. MUD helps to reduce the overall threat vector surface. ACLs [92] are normally defined using classes, these are based on the MAC and IP address when deployed on the network switches. MUD working principle is as: the device is allocated with the MUD URL. This MUD URL is used for accessing the MUD ACL file. This service is given by the MUD server, MUD server fetches the MUD ACL file from the vendor or manufacturer website. Then install this ACL file for this particular device in the switches. Only after this authentication process, the device is allowed to communicate over the network.

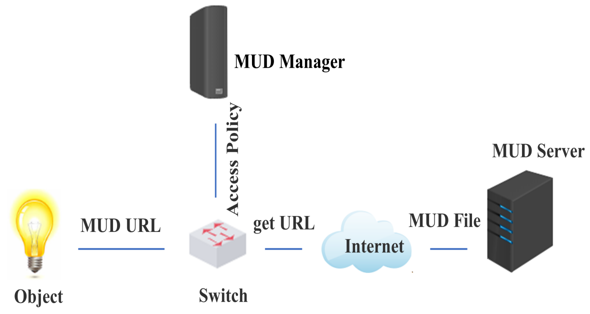


Fig. 9. MUD Working Principal

IV. Application of SDN, ML and MUD in IDPS of IoT

A. Detection using SDN based on Anomaly and Entropy

Attack detection is a crucial aspect of network management. The SDN feature of global network visibility plays a vital rule in network security. This feature helps in monitoring network real-time status. This helps entropy-based detection techniques to become more effective in SDN networks. SDN uses a set of protocols to communicate between the application layer and the controller and vice versa, such as RESTapi, XML and netconf, etc. also controller uses certain communication protocols to talk to the data layer and vice versa like sflow and OpenFlow, etc.

Anomaly Detection covers range of detection systems, as shown in the study [93] communication protocol OpenFlow and sampled flow (sflow) can be used for anomaly detection. The study uses threshold random walk with a connection-based algorithm for the entropy-based anomaly detection. The study also compares the flows required for anomaly detection in the case of sflow and OpenFlow. For sflow the number is 217 and for OpenFlow, it is 5184. This concludes that the sflow is better than OpenFlow in the context of traffic collection for anomaly detection. Another study [94] uses a two-stage scheme for the detection of DDoS attacks. The study first set some threshold value for the flows after that is if some flow crosses the limit it is sent to the detection stage. As in the experiments the author set 3000 packets for five seconds is allowed if some flow exceeds the limit then the 800 packets per second (PPS) for 5 s is imposed. Also, the packet filtering mechanism is activated. One of the research [95] implement anomaly detection algorithms at the border router of the home network. The study uses the NOX SDN controller. The algorithms were also based on threshold random walk with credit-based, rate limiting, maximum entropy detector, and NETAD. And the results show that the algorithms perform better in small networks like home networks.

Another study [96] developed a bidirectional sketch algorithm to detect attacks. It identifies the IP address of the destination for the asymmetric traffic pattern depends on the threshold value to detect some anomaly. once the malicious traffic is identified the controller instructs the switch to drop the malicious flows. In this, once the malicious traffic is identified it is detected and blocked at the ingress points of the network to avoid collateral damage. The author [97] present a distributed algorithm for anomaly detection based on the entropy technique.

Entropy detection is one way to detects attacks, as the study suggests that most of the entropy detection is based on the collection of flows from the network, so if the network is big then due to a large number of flows may overload the network. Therefore, the study processes the flows at the switch and present filters for the DDoS attack at the edge switches. The study in [98] introduces the NetFuse device between the switch and the controller to monitor the network load. This instructs the switches to reroute the flows causing the congestion. StateSec [99] a novel algorithm to detect the DDoS attacks using the port scans. Anomaly is measured by the abrupt changes in the traffic features. However, the study shows no implementation for this algorithm. Few industrial solutions have been proposed Radware [100] developed DefenseFlow to detect the DDoS attacks. This technique measures some key attributes like packet rate, average packet size, bandwidth connection distribution, and connection rate. In case of attack, the device instructs the controller to reroute the traffic to the specialized devices to handle the traffic.

B. Detection using ML

ML techniques in SDN networks help against the detection of the malicious flow as in the study [101] detect the anomalous flows in the SDN network. The information is taken from the flow tables from the switches and gets the flow of information. The DPTCM-KNN algorithm in the detection module process these anomalous flows. The only issue here is the processing overhead as the process is repeated after 10 seconds. In another study proposed a trust-based approach for the detection of malicious devices using the packet data and device profile.

Another study [102] proposed DDoS detection using OpenFlow by proposing the self-organizing maps (SOM) for attack detection. The SOM is an unsupervised neural network method for classifying the traffic as normal or abnormal. In this method, the controller continuously collects the data from the switches and other devices and monitor the attributes like average packets in the flow, average number of bytes in the flows, and the average duration of flows. The data is then fed into the classifier for the detection of DDoS attacks. All the process also adds some overhead on the processing of the controller. One of the studies [103] analyzes the ML algorithms in the context of the SDN for providing security, resource optimization, traffic classification, and quality of service. This further shows that ML brings intelligence to the controller for the detection and prevention of attacks.

The study [104] shows that ML can be used for DDoS and intrusion detection also shows the pros and cons of the surveyed mechanism in handling the intrusion detection in SDN networks. Another study [105] analyzes the traffic scheduling problem in the SDN network in a hybrid data centre environment. In this, the edge devices are responsible for the ML-based elephant flow traffic classification. This reduces the burden on the SDN controller. However, the scope is limited it is unable to give the network-wide information having multiple switches and controllers.

A study EUNOIA [106] proposes a system to detect the threat and respond in the SDN system. The framework has multiple modules like data preprocessing, decision making, response system, and predictive data modelling. For predictive data modelling, the system uses the decision tree and random forest algorithms, to monitor malicious and suspicious traffic. Response module works based on alerts. While the decision-making module is responsible for the routing path computation and implements new flow rules for each flow type. The experimental results display high detection accuracy and reduce data preprocessing time. However, if the data is increased from any node then the processing time also jumps up showing degradation in overall performance also the scalability is an issue in this framework.

Another study presents a management framework for classification, anomaly detection, and mitigation within the SDN network. The system works in two phases for classification and detection. First is the lightweight phase in this light computation is done to detect any suspicious or malicious activity. In the second phase, SVM, an ML algorithm, is applied for the classification and abnormal flow detection. Again, the controller requires a lot of processing power to poll all the switches in the network. In [107], propose a scheme using the DL algorithm for the detection of DDoS within an SDN network. The study develops stack autoencoder using the SoftMax classifier and unsupervised deep learning algorithm. The results based on the NSL-KDD dataset shows around f-score 75.76% and 69% accuracy. Again, this technique requires to monitor each packet that limits the controller performance.

A similar study [108] uses deep neural networks in the SDN network for intrusion detection. The author developed the neural network. The network constitutes one input, three hidden, and one output layer. The SDN controller receives data from the switches in the network and sends the data to the detection module. The results using the NSL-KDD data set is about 75.75%. The scalability of SDN has been tried to handle in [109] by proposing the framework called Athena. The main concept of this framework is to implement the detection mechanism not only on the edges but to deploy this at network-wide switches and controllers for better results. Athena doesn't require special hardware for the deployment of the detection modules. But the framework doesn't provide adaptive measurement for resource optimization.

C. Detection using MUD

The researchers provide IoT security solutions using MUD as in [110] implement MUD over the OpenFlow switches in the SDN network. In another study, [111] they developed a machine for the detection of anomalous patterns in a MUD compliant network. Also analyzing the IoT network behaviour after volumetric attacks happen in the network. When they compare the results for attack detection with other systems its detection is much superior. Similarly, in [112] translated the MUD policies into the flow rules and implement these policies using SDN technology. The study also showed the limitations of MUD based policies in the context of securing the network. The author in [113] sorted out the validation and integration problems of the MUD profiles with the network. Further, the study validates the MUD profile for each device also makes sure that the profile is confirmed under the organization policy.

D. Prevention using SDN and ML

The ability of the SDN controller to be programmed and change the rules on the fly made it suitable for attack mitigation. As the security notification can be shared across the network. Based on the context flow rules that can be generated and implemented on the edge devices. Also, the anomaly detection capability of the ML technique made it possible to develop the defence system based on the collaboration of ML and SDN. One of the study Drawbridge [114] proposes a framework for ISP used for attack mitigation. The main objective of the framework is to stop the dropping of the customer traffic from the ISP due to heavy load. The detection in the framework is performed at the end devices like switches. The controller and the switches in the ISP share the rules. The responsibility of the validity check of the rules lies on the controller before deploying them on the switches.

In another study proposes the SENSS [115] that offer an interface for the attack mitigation. When the attack happens, the victim sends the information to the controlling body, which is ISP in this case, with all the routes and traffic details. However, the system allows the victim to request for the rerouting the network traffic. In this study, Bohatei [116] the author leverages the strength of network function virtualization (NFV) along with SDN to detect the DDoS attacks. The authors make use of NFV technology to place and start the defence virtual machine at the required location in the network. This developed architecture pushes all the network traffic at these initiated VMs. This framework works for the ISP network. The kind of architecture of ISP provides support to create a service for the customers to defend them from DDoS attacks. After detecting the anomaly in the network another process of estimation starts on the suspicious traffic. This estimation is sent to the resource manager module to find out the number, type, and location for the instantiation of virtual machines. The process is based on a couple of algorithms namely server selection and data centre algorithm used in the data centre. The results show that the system can mitigate the DDoS attacks in one minute. But this put quite a workload on the ISP as it must handle hundreds and thousands of customers.

In other research, [117] a framework is proposed for collaborated defending against the DDoS attacks. In this, the customer requests to ISP to provide the service against the DDoS attack. Upon request, the suspicious flows are sent to the middleboxes for further processing. But the implementation of the said framework has not been done. One of the frameworks proposed called ArOMA [118] it tries to mitigate the DDoS attacks automatically using the SDN strengths like centralized manageability and programmability at the ISP. This framework brings all the three stages like monitoring, detection, and mitigation against DDoS attacks under one automated umbrella. This framework also enables the collaboration between the customers and ISP to defend against DDoS attacks. The author provided the implementation and evaluation of the proposed framework.

In the research, [119] presented a framework based on the collaboration of SDN and content-based data network (CDN) to manage the high volume video traffic flows. Normally this application is deployed at the controller to get the hidden knowledge of the network like topology end-users to optimize the network. The ISP uses the application to manage the huge traffic emanating from CDN. The CAPTCHA [120] is used for the protection of the services. The server runs these services. The results in this process are based on weak programming of the bots. Also, they believe that the bots cannot perform the IP spoofing which is not the case.

E. IDPS Systems for IoT based on SDN, ML and MUD

IDPS has become a very eye-catching technique especially with the involvement of a new state of art techniques. Fig. 10 shows the integration usage of MUD, SDN, and ML in IDPS of IoT [111]. IoT devices have their local network in which they communicate with each other and also communicate with the Internet through a gateway. The SDN switch manages the flow-table rules dynamically. The MUD engine works with the SDN controller and App. The architecture also contains the MUD collector along with an anomaly detector and IDPS. These elements work combinely to manage the flow-table rules embedded in the switch and also monitor the device activities inside the network [112].

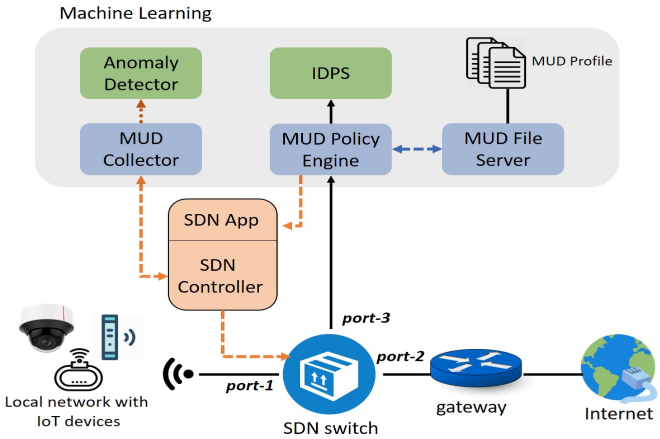


Fig. 10. Network Architecture of IoT based on SDN, MUD and ML (redrawn from [111])

In this network architecture, default rules are initially configured in the SDN switch to mirror packets that are intended to use for the device identity. MAC addresses are generally used for the identification of the device. Application for example DHCP contains this MAC of the device and provides MUD-URL which adapts the MUD standard. This is used by the MUD policy engine to discover new IoT devices making a connection with that network. The MUD engine already has a record of discovered devices. The MUD engine retrieves the MUD profile from a MUD file server and stores it till its validity. To be noted here that the manufacturer operates and manages the MUD file server. The MUD policy engine makes the ACLs (access control lists) of the MUD profile into the set of flow rules. The packets intended to be sent are mirrored and send to the IDPS for the inspection. IDPS confirms the traffic that does not follow the MUD specification. Some traffic still can pass to the network by using spoofing techniques. Therefore, an anomaly detector has been used. For these, MUD collector pulls flow counters from the SDN switch, then compute each device's attributes, and stream them to that anomaly detector. ML is used to learn the policy, anomaly detection, and flow rules for the smooth operations of the whole architecture. ML is also used to train the system to determine the attack flow. The system is also trained by using ML to detect the abnormality of the expected traffic (defined by the MUD profile). This helps the system to detect attacks.

These technologies like SDN, ML, and MUD are getting attention from the researchers and industrialists to be used in intrusion detection systems. This improves the overall performance of the system and able them to meet future challenges. As in [143], the study proposes an IDPS system for the protection of the command and control system for unmanned vehicles (UV). There are attacks like anti-drone and anti-autonomy on these UVs and these attacks disturb their autonomous decision-making system. The system uses a hybrid IDS system approach. Another study [144] more focused on wireless sensor networks shows the need for IDPS system to protect the network from growing cyber-attacks. Further, the author compares the ML and DL approaches in the context of the IDPS. Another paper [121] proposes a secure model for smart cities based on the IDPS and DL approaches. They propose a deep migration learning model and use KDD-CUP 99 data set for the model evaluation. The results show shorter detection time and higher detection efficiency. One of the studies [122] proposes a security model for the IoT devices based on the strength of the SDN network. It combines the power of firewall and IPS to detect anomaly in the network traffic and detect the attack in the IoT network.

A new concept known as Manufacturer Usage Description (MUD) has been used with SDN [111] for intrusion detection. MUD is a device description for its expected behaviour, provided by the vendor. The system based on SDN, MUD, and ML approaches to defend against benign and volumetric attacks in IoT devices. In this study, [112] the author implements the MUD in the SDN network and analyzes the effect of this new technology against the volumetric attacks in IoT networks. A similar work [113] generates MUD profiles based on the behaviour of the device, the work also validates the generated profiles with the organizational policies. Another study [110] presents the MUD implementation in SDN based network. It also shows the implementation scenario in a scalable fashion. An interesting study [123] shows that the MUD is not completely defining the IoT devices, so they propose their security framework with few extensions in a MUD. The author presents their learning model that extracts the IoT device feature by analyzing the network traffic. Based on these features develop a normal behaviour profile of the device. Thus, the claim that the specification of the device using this framework is tighter and clearer.

TABLE IV. Taxonomy of IDPS systems for IoT based on technology

References	SDN	ML	MUD/ Device profile	Technique	Tools	Remarks
Li, et al. [121]	✗	✓	✗	Migration Learning	KDD CUP 99 Ubuntu MATLAB	The IDPS does not consider the resource limitation of the IoT devices. Also, the vulnerabilities of the IoT devices itself.
Gonçalves, et al. [122]	✓	✗	✗	NIDS	Snort Linux Nmap OVS	Use SNORT IDS for detection and prevention is done by SDN. Limiting the system capability for the detection of a few attacks.
Hamza [111]	✓	✓	✓	MUD SDN ML	OpenFlow switch Faucet SDN controller MUD policy engine	Rely on MUD for IoT security. But MUD has its limitation for completely defining the IoT device profile.
Hamza [112]	✓	✗	✓	MUD SDN	MUDgee PCAP	Only focus on MUD implementation rather than the comprehensive security solution for IoT.
Hamza [113]	✗	✗	✓	MUD	MUDgee	Only provide proof of concept for the MUD profile generation and did not focus on IoT security.
Ranganathan [110]	✓	✗	✓	SoftMUD	ODL YANG	The security solution is only able to test the DDoS. Attacks like malware are real threats to IoT devices.
Singh, et al. [123]	✗	✗	✓	MUD Clustering Technique	Multitech Conduit LoRa Gateway Ettus USRP B210 Linux containers LXD TShark	Only focus on removing the weakness of the MUD. Did not address how to handle the diverse cyber-attacks on IoT devices.
Kumar and Lim [124]	✗	✓	✗	Random Forest k-NN Gaussian Naive Bayes	scikit-learn Wireshark	Due to the absence of the MUD like technique the solution capability to secure diverse IoT devices is a question mark.
Amangele, et al. [125]	✓	✓	✗	decision-tree LR LDA KNN CART NB SVC	Scikit-learn Python package CICIDS2017 Dataset	In this solution, scalability will be an issue. As the number of IoT devices increases there is more processing load on the SDN controller that effects the real-time detection of the system against intrusion detection.
Wani and Revathi [126]	✓	✓	✗	BPNN	NSL-KDD Dataset RYU Controller	The proposed solution has not been tested in a real-time environment. Also, only a flood attack has been tested, the system performance for another kind of DDoS attacks is not tested.
Chang [127]	✓	✓	✓	Self-Organizing Map (SOM)	YANG data model Mininet Ryu SDN controller	The proposed solution results are not compared with any other existing detection and prevention system.
Wu, et al. [128]	✗	✓	✓	Signature based Device Profile	NS-3	Effective only against the EEA attacks
Venkatraman and Surendiran [129]	✗	✓	✗	Automata Technique	Automata controller	The IDS system is designed for the home security systems
Soe, et al. [130]	✗	✓	✗	ML Correlated-Set Thresholding on gain-ratio (CST-GR) algorithm	Raspberry Pi system Botnet DataSet	Only works for the known attacks.
Putra, et [131] al	✗	✓	✗	SVM, Block Chain	Raspberry Pi system	This IDS make additional overhead of blockchain
Li, et al. [132]	✗	✓	✗	DL	NSL-KDD Keras	The system has limitation for the new cyber-attacks
Ferrag, et al.	✗	✓	✗	ML	CICIDS2017 BoT-IoT	Shows little effectiveness on some attacks
Ferrag, et [133] al.	✗	✓	✗	specification heuristic	CUPCORBAN JAVA Platform	It adds to the overhead for the IoT devices
Nguyen, et [134] al	✗	✓	✓	federated learning approach	Kali Linux tcpdump	The System just talk about the detection but didn't give much about prevention from such attacks
Cervantes, et al [135]	✗	✓	✓	Behavior Based	Cooja	The study focuses only on sinkhole attack. Need to test against new attacks types
Eskandari, et al. [136]	✗	✓	✓	Behavior Based	Raspberry Pi 3 model B AGILE gateway software	The work does not describe the overhead and the performance plenty for the IDS system used in the IoT based networks
Bhale, et al. [137]	✗	✗	✓	Device profile	Contiki OS Coja Simulator	The IDS system just focuses on one type of attack sinkhole attack. Its role against other attacks need to be tested.
Babu and Reddy [138]	✗	✓	✓	Specification based	CUPCORBAN JAVA Platform	The IDS can lead to more false positive as compare to signature based technique
Ambili and Jose	✗	✗	✓	Behavior Based Blockchain	IoT Devices	The performance of the IDS system is not very commendable
Al-Duwairi, et al [139]	✗	✗	✗	SIEM	Splunk SIEM	The IDS system is not compared with other IDS in terms of performance
Mudgerikar, et al [140]	✗	✓	✓	Anomaly base Profile Base	IoT Device	This is system level IDS tailored for IoT devices. But its comparison with network IDS system is not mentioned.
Kumar, et al. [141]	✗	✓	✗	ML Decision Tree	UNSW-NB15 data set	Limited for other kind of attacks like zero day attacks.
Breitenbacher, et al. [142]	✗	✗	✓	host based	Linux kernel	The system overhead has not been measured

EDIMA [124] as an early detection system for the IoT devices. Thus, the system helps in detecting malware in the network, especially for large-scale networks. The system uses an ML algorithm at the edge devices for traffic classification. One of the studies [125] proposes a hierarchical ML approach in the SDN network to reduce processing load at the edge devices for anomaly detection. The results show that the two-level classifier significantly reduce the packet processing at the edge. The research [145] shows the pons and cons of different IDS and IPS systems in the context of cloud computing. The author in [126] presents an IDS system SDIoT-IDS based on SDN for the IoT devices such that the maximum load is taken off from the edge devices. The system is tested only with few attacks like flood attack and ICMP based attacks. Table IV summaries the IDPS systems discussed with addition to the classification of systems based on the technology used in them.

1) Role of SDN, ML AND MUD in IDPS: The IDPS system based on the MUD, ML and SDN become more robust and dynamic. The MUD provides the system first line of defence by verifying the ID and the role of the device within the network. This reduces the threat surface for the IoT devices. The second line of defence, the ML/DL techniques helps not only to detect the established cyber threats but also the unknown threats that can cause the system to malfunction. Finally, the last line of defence is the SDN technology that helps in taking realtime actions to rectify the damage caused by the threats and prevent them to happen again by enforcing new rules and policies within the network. All these layers are shown in the Fig. 11.

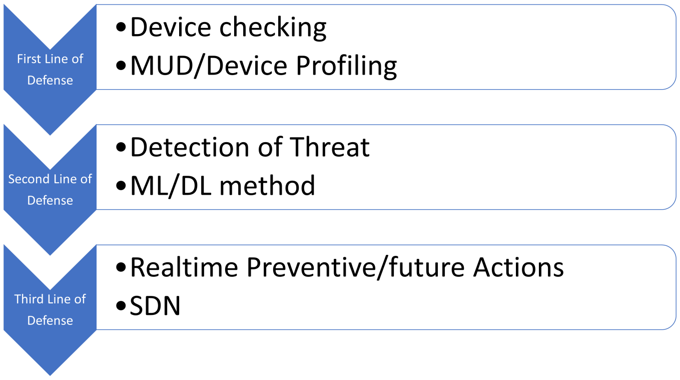


Fig. 11. Role of MUD, ML and SDN in IDPS

2) Performance of IDPS in IoT: Research [127] shows the advantages of using MUD and SDN together against the flood attack in IoT. The results show 98% detection for TCP attacks and 90% detection for UDP attacks. With the introduction of the IoT device profile for the IDPS systems will reduce the attack surface. The [128] study used graph theory to describe the network. They used centralized and malicious node detection (CAMD) IDS. It uses the genetic algorithm to analyze the nodes data gathering behaviour. This node profile helps for the distribution and passive EEA resistance (DPER) module used as the second part of IDS to lower the Energy Exhaustion attack (EEA). The accuracy for detection is 100%. Also in [142] the author purpose HADES-IoT is a host-based IDS system for the IoT devices. This is a lightweight IDS system that requires very few resources. It defends the IoT devices using profiling techniques. Profiling is performed on every IoT device to detect malware like VPNFilter, Persirai, Marai, and IoTReaper. The results are 100% detection. As there is a serious need for the including MUD in IDPS systems for the end to end security of the IoT devices.

Little work has been on the MUD and its implementation for IoT as the standard is relatively new as compared to other technologies. However, the ML and SDN usage in IDPS systems are showing convincing results. In this research, [130] a lightweight ML-based IDS system has been proposed for the IoT devices. In this, a novel feature selection algorithm has been implemented known as correlated-set thresholding on gain-ratio (CST-GR) algorithm. Giving the detection accuracy of 99.4%. Similarly in [133], the study proposes an intrusion detection system based on tree-based and rule-based classifiers. It uses EP Tree, JRip algorithm, and Forest PA. It shows good performance in the context of accuracy, false alarm, detection rate, and false alarm. The detection is some times vary between 30% to 100% depend on the attack type. Table V summaries the IDPS systems performance based on the detection of specific cyberattacks.

The performance of the IDPS system as shown in Table V is very interesting as few systems show the 100% detection accuracy. As we can see that [128] shows 100% detection accuracy, similarly, in [142] the detection rate is 100% for the latest attacks. However the minimum detection in the current sample is close to 80% depend on the attack type as shown in [140], which is still quite encouraging. Keeping the fact that using the new technologies in the IDPS system made them more dynamic and up to date to handle the upcoming cyber threats, as compared to the conventional techniques that require predefined rules and signatures to identify the threat. The range and application of the IDPS systems were limited. Also, the response to those threats was not real-time and it took long to rectify the system and update it to counter such attacks. As the technology gets more mature the IDPS systems will get more effective and applicable to provide end to end security for IoT networks.

V. Open Issues, Challenges and Future Research Directions

In the previous sections, we have reviewed some aspects of the combined use of SDN, ML, and MUD in IDPS of IoT. Together these techniques provide several common research issues in different aspects of IDPS of IoT. In this section, we highlight some of the future research issues and challenges in this domain.

A. End to End Security for IoT using IDPS and SIEM

All the types of IDPS discussed above for the security of the IoT has some drawbacks too. Like network-based, NBA, wireless, and host-based all have different characteristics. Each of the technology can detect a set of attacks that the other cannot. So, if we use multiple types of IDPS together it will produce much better results against the malicious activity. Also, as each IDPS works independently so there is no fear of a single point of failure. But if all the IDPS systems are not integrated then their effectiveness is limited.

TABLE V. Performance of IDPS systems for IoT based on technology

References	Attacks	Performance
Li, et al. [121]	DoS Probe	97.2%
Gonçalves, et al. [122]	Port Scanning DoS	—
Hamza [111]	port scanning, TCP/UDP/ICMP flooding ARP spoofing TCP/SSDP/SNMP reflection	92.82%
Hamza [113]	DoS	—
Ranganathan [110]	DDoS	—
Singh, et al. [123]	Malicious Nodes	—
Kumar and Lim [124]	Mirai, Hajime, Remaiten, Linux.Wifatch Brickerbot, Satori, Masuta, Linux.Darlloz, Reaper, Amnesia	94.44%
Amangele, et al. [125]	Slowloris, Hulk, Golden Eye, Heartbleed Slowhttptest , DDOS, BOT	99%
Wani and Revathi [126]	TCP flood ICMP Attack	99%
Chang [127]	TCP SYN Attacks UDP Attacks	90%
Wu, et al. [128]	Energy Exhaustion attack (EEA)	100%
Venkatraman and Surendiran [129]	DoS Hijacking attacks, Zero day attacks Replay attacks	99.06%
Soe, et al. [130]	DDoS attacks, probing attacks (reconnaissance), information theft attack	99.4%
Putra, et al [131]	DoS, Man-in-the-middle, Spoofing Reconnaissance, Replay attacks	—
Li, et al. [132]	DoS, Probe, R2L, U2R	82.62%
Ferrag, et al. [133]	DOS, Brute, force, web, Attack, SQL injection	30%-100% depend on the attack type
Babu and Reddy [138]	malicious traffic	91%
Nguyen, et [134] al	malware , attacks	95.6%
Cervantes, et al [135]	Sinkhole attack	Mobile Devices 75% Fixed Devices 92%
Eskandari, et al. [136]	Port, Scanning, HTTP and SSH Brute Force, SYN Flood attacks	98%
Bhale, et al. [137]	Sinkhole attacks	TPR 95.86% TNR 94.31%
Babu and Reddy [138]	malicious traffic	91%
Al-Duwairi, et al [139]	botnet, Attacks, Flood Attack	—
Mudgerikar, et al [140]	malware, attacks, brute-forcing, DDoS, crypto-mining	78% - 99%
Kumar, et al. [141]	DoS, attacks, Probe, attacks	88.92%
Breitenbacher, et al. [142]	VPNFilter, IoTReaper	100%

The IDPS can be directly or indirectly integrated for overall improved performance. Direct integration involves using multiple IDPS from the same vendor. This integration shows quite an improvement in the accuracy of the system to detect the threats. But this approach has a flaw if any single module failed the whole system will be compromised. To address this issue an indirect integration approach is used to achieve the same result. In this scheme, all types of the IDPS report to one system called security information and event management (SIEM) software [146].

SIEM [147] can be used to complement to IDPS system. This system can analyze the data from different IDPS and detect the attack more efficiently. To protect the future-critical applications based on IoT requires an end to end security management system. This paradigm requires to orchestrate and manage security across all the connected domains, like connected devices, networks of communication, cloud, apps up to the users. This is only possible by using existing Intrusion Detection and Prevention Systems (IDPS) and Security Incident and Event Management (SIEM).

B. Next Generation Firewall over IDPS

With the advent of ML techniques in IDS and IPS, the predictive mechanism becomes very efficient in changing rules based on previous history. This model helps to detect the unseen attacks. But the scheme fails in the generative policy model, as IDS and IPS generate new policies based on the context. This scenario becomes a challenge for automated and reliable policy-based management systems. Especially in distributed and coordinated systems.

Firewalls in this case use packet information and directly block the suspicious packets rather than to detect and block the attack. But IDS and IPS can handle many unseen attacks that the firewall cannot detect. But at the cost of time and resources. That is not acceptable to the IoT devices having a limited resource. Therefore, a next-generation firewall using ML and SDN approaches to analyze the IP packets for malicious packet detection and block unseen attacks rite at the edge before entering into the network will be more efficient and suitable for the IoT based networks [148].

C. Lack of Intrusion Detection Dataset

The available intrusion detection dataset is not up to the mark for the research predictions as the academic's research require proper classification of data. For this reason, the network researcher uses artificial datasets for the network intrusion detection because of the unavailability of the realistic datasets. So, the importance of realistic datasets cannot be undermined for the accurate and more realistic testing and evaluation of the intrusion detection systems. The most common dataset used nowadays are KDD cup 1999 with a new version and NSL-KDD for network-based intrusion detection system [13].

D. Gaps in Machine Learning, SDN and MUD

Deep learning is expected to improve the security of the network, but infect it is also vulnerable to cyber-attacks. As if the input to the DL algorithms is changed the output of the system could be drastically different [149]. As an example, the use of DL in IDPS if the input to these algorithms is changed then the IDPS system will be in the control of the adversaries. Therefore, more investigation and care are required while using deep learning algorithms in critical security applications. One of the issues with the SDN networks is the logically centralized controller. In this context, the attacks on the controller pose a more serious threat over all the networks.

Moreover, the communication channel between the controller and the switch can also be compromised. There are other challenges involved with SDN is scalability and compatibility with the conventional networks [150]. Similarly, MUD also has few weaknesses, as shown in [112], when the MUD translated to the flow rules it has vulnerabilities against the internal and external attacks on the IoT devices. This MUD profile needs to add on a few additional attributes regarding the devices for shortening the attack surface.

E. Usage of Blockchain

Blockchain has revolutionized the cryptocurrency industry. The main design is consisting of the secure distributed database (a.k.a public ledger) all the participants do their transactions from this. The cryptocurrency like Bitcoin and Ethereum do their transactions in peer to peer architecture. The working principle of blockchain is such that when one peer wants to make a transaction to another peer it makes transaction requests to all the peers in the blockchain network. This way every node gets periodic transaction updates and puts them into a single block. After that, the validation of each block is done by a special consensus algorithm that is executed by special nodes in the network known as miners.

The new IoT emerging applications can take advantage of the secure communication architecture of blockchain [151]. Since in the IoT world more and more devices and sensors need to communicate to provide real-world applications, Blockchain gave a tamper-resistant record allowing the participants to have secure and consistent access. Additionally, the blockchain provides flow management. It provides an efficient way to automate the business and creating smart contracts [152]. IoT takes benefit from blockchain due to the certain feature of the technology;

Decentralization: Due to the decentralized architecture of IoT blockchain suits as a security solution. The default decentralized architecture of IoT solves the problem of single failure and at the same time gives robustness against DoS attacks. Pseudonymity the public keys are the identifications of the nodes in the blockchain. But these Pseudonyms don't give any information regarding the identity of the node.

Secure Transaction: every transaction send is first signed by the node itself, then it gets verified and validated by the miners. Once this transaction got verified it cannot be altered and whole proof of the traceable events is stored in the network.

F. Software Watermarking

Watermarking is a technique used for software protection from the cyber-attacks [153]. As the algorithms use keywords within the code to produce a key. Further, there are extraction algorithms that extract the key and compare it with the original key to check the amount of tempering in the software. if the tampering reaches to certain threshold the intrusion detection system can take certain action to eliminate the malicious software or block it from further execution. This technique can be used in the intrusion detection and protection systems to enhance the overall system security and reduce the attack surface, especially when used with MUD.

G. Hardware Limitation

The use of IDPS including ML, MUD, and SDN, we need a more sophisticated computational capability, more memory resource, energy or power, etc. However, the smaller IoT devices still have a resource constraint. The IoT network requires more computational power and better network infrastructure to fully benefit from the applications of these latest technologies. Further, the requirements for real-life implementation and real-time monitoring are more resource-demanding approaches in the context of IoT network security. Therefore, more research needs to be performed in the hardware part to include the integrated application of SDN, ML, and MUD for the IDPS in IoT.

VI. Conclusion

This paper discusses the security issues of the Internet of things (IoT) and the role of the Intrusion Detection and Prevention System (IDPS) to address these challenges. These IDPS are not new in network security, whereby many conventional IDPS systems are already being used for security purposes. However, in this study, we focus on IDPS systems based on modern and state of art technologies like Software Defined Network (SDN), Machine Learning (ML) and Manufacturer Usage Development (MUD) techniques. Further, we analyze the application and the effects of the latest IDPS systems on the security of the IoT networks and devices. The new concept of security design is evolving among the researcher communities and industries to provide comprehensive security in IoT networks. MUD is one of the latest standards developed in this direction. This study also analyzes the integration of MUD with IDPS systems to study their effect on network security. For the future of IoT networks, the importance of the IDPS system is irrefutable. Systems like security information and event management (SIEM) and IDPS using the power of modern technology are the future of security in IoT networks.

Acknowledgment

The authors would like to thank the sponsor as this work was supported by the University of Malaya, Malaysia (Grant No. GPF017D-2019).

References

[1] P. I. Radoglou Grammatikis, P. G. Sarigiannidis, and I. D. Moscholios, “Securing the internet of things: Challenges, threats and solutions,” *Internet of Things*, vol. 5, pp. 41–70, 2019.

[2] K. Kalkan and S. Zeadally, “Securing internet of things (iot) with software defined networking (sdn),” *IEEE Communications Magazine*, no. 99, pp. 1–7, 2017.

[3] O. Flauzac, C. González, A. Hachani, and F. Nolot, “Sdn based architecture for iot and improvement of the security,” in *2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops*. IEEE, 2015, Conference Proceedings, pp. 688–693.

[4] “Rfc-8519 - proposed standard mud,” 2019. [Online]. Available: https://datatracker.ietf.org/doc/rfc8519/

[5] S. Gangadhar and J. P. Sterbenz, “Machine learning aided traffic tolerance to improve resilience for software defined networks,” in *2017 9th International Workshop on Resilient Networks Design and Modeling (RNDM)*. IEEE, 2017, Conference Proceedings, pp. 1–7.

[6] S. Nanda, F. Zafari, C. DeCusatis, E. Wedaa, and B. Yang, “Predicting network attack patterns in sdn using machine learning approach,” in *2016 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*. IEEE, 2016, Conference Proceedings, pp. 167–172.

[7] N. A. Azeez, T. M. Bada, S. Misra, A. Adewumi, C. Van der Vyver, and R. Ahuja, *Intrusion Detection and Prevention Systems: An Updated Review*. Springer, 2020, pp. 685–696.

[8] M. T. S. P. Tiwari and P. Mishra, “Review of intrusion detection system,” *International Journal of Scientific Research & Engineering Trends*, 2019.

[9] T. Salman and R. Jain, “A survey of protocols and standards for internet of things,” *arXiv preprint arXiv:1903.11549*, 2019.

[10] K. A. P. da Costa, J. P. Papa, C. O. Lisboa, R. Munoz, and V. H. C. de Albuquerque, “Internet of things: A survey on machine learning-based intrusion detection approaches,” *Computer Networks*, vol. 151, pp. 147–157, 2019.

[11] S. Choudhary and N. Kesswani, “A survey: Intrusion detection techniques for internet of things,” *International Journal of Information Security and Privacy (IJISP)*, vol. 13, no. 1, pp. 86–105, 2019.

[12] S. Hajiheidari, K. Wakil, M. Badri, and N. J. Navimipour, “Intrusion detection systems in the internet of things: A comprehensive investigation,” *Computer Networks*, vol. 160, pp. 165–191, 2019.

[13] N. Sultana, N. Chilamkurti, W. Peng, and R. Alhadad, “Survey on sdn based network intrusion detection system using machine learning approaches,” *Peer-to-Peer Networking and Applications*, vol. 12, no. 2, pp. 493–501, 2019.

[14] R. Sahay, W. Meng, and C. D. Jensen, “The application of software defined networking on securing computer networks: A survey,” *Journal of Network and Computer Applications*, vol. 131, pp. 89–108, 2019.

[15] P. B. Pajila and E. G. Julie, “Detection of ddos attack using sdn in iot: A survey,” in *Intelligent Communication Technologies and Virtual Mobile Networks*. Springer, 2019, Conference Proceedings, pp. 438–452.

[16] F. Restuccia, S. D’Oro, and T. Melodia, “Securing the internet of things in the age of machine learning and software-defined networking,” *Ieee Internet of Things Journal*, vol. 5, no. 6, pp. 4829–4842, 2018.

[17] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, “Internet of things: A survey on enabling technologies, protocols, and applications,” *IEEE communications surveys and tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.

[18] M. Sharma and S. C. Gupta, “An internet of things based smart surveillance and monitoring system using arduino,” in *International Conference on Advances in Computing and Communication Engineering (ICACCE)*. IEEE, 2018, Conference Proceedings, pp. 428–433.

[19] G. Coley, “Beaglebone black system reference manual,” *Texas Instruments, Dallas*, vol. 5, 2013.

[20] U. Isikdag, *Internet of Things: Single-board computers*. Springer, 2015, pp. 43–53.

[21] D. K. Rahmatullah, S. M. Nasution, and F. Azmi, “Implementation of low interaction web server honeypot using cubieboard,” in *International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC)*. IEEE, 2016, Conference Proceedings, pp. 127–131.

[22] G. Chu, N. Apthorpe, and N. Feamster, “Security and privacy analyses of internet of things children’s toys,” *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 978–985, 2018.

[23] A. Khattab, Z. Jeddi, E. Amini, and M. Bayoumi, *RFID security threats and basic solutions*. Springer, 2017, pp. 27–41.

[24] F. Li, A. Shinde, Y. Shi, J. Ye, X.-Y. Li, and W. Z. Song, “System statistics learning-based iot security: Feasibility and suitability,” *IEEE Internet of Things Journal*, 2019.

[25] “The iot attack surface: Threats and security solutions,” 2019. [Online]. Available: https://www.trendmicro.com/vinfo/gb/security/news/internet-of-things/the-iot-attack-surface-threats-and-security-solutions

[26] N. A. Alrajeh, S. Khan, and B. Shams, “Intrusion detection systems in wireless sensor networks: a review,” *International Journal of Distributed Sensor Networks*, vol. 9, no. 5, p. 167575, 2013.

[27] M. R. Thakur and S. Sanyal, “A multi-dimensional approach towards intrusion detection system,” *arXiv preprint arXiv:1205.2340*, 2012.

[28] B. B. Zarpelão, R. S. Miani, C. T. Kawakani, and S. C. de Alvarenga, “A survey of intrusion detection in internet of things,” *Journal of Network and Computer Applications*, vol. 84, pp. 25–37, 2017.

[29] F. Li, Y. Shi, A. Shinde, J. Ye, and W. Z. Song, “Enhanced cyber-physical security in internet of things through energy auditing,” *IEEE Internet of Things Journal*, 2019.

[30] H. Sedjelmaci, S. M. Senouci, and T. Taleb, “An accurate security game for low-resource iot devices,” *IEEE Transactions on Vehicular Technology*, vol. 66, no. 10, pp. 9381–9393, 2017.

[31] J. Li, Z. Zhao, R. Li, and H. Zhang, “Ai-based two-stage intrusion detection for software defined iot networks,” *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2093–2102, 2018.

[32] K. Scarfone and P. Mell, “Guide to intrusion detection and prevention systems (idps),” National Institute of Standards and Technology, Report, 2012.

[33] K. Letou, D. Devi, and Y. J. Singh, “Host-based intrusion detection and prevention system (hidps),” *International Journal of Computer Applications*, vol. 975, p. 8887, 2013.

[34] J. P. Amaral, L. M. Oliveira, J. J. Rodrigues, G. Han, and L. Shu, "Policy and network-based intrusion detection system for ipv6-enabled wireless sensor networks," in *IEEE International Conference on Communications (ICC)*. IEEE, 2014, Conference Proceedings, pp. 1796–1801.

[35] J. R. Vacca, *Computer and information security handbook*. Newnes, 2012.

[36] H.-J. Liao, C.-H. R. Lin, Y.-C. Lin, and K.-Y. Tung, "Intrusion detection system: A comprehensive review," *Journal of Network and Computer Applications*, vol. 36, no. 1, pp. 16–24, 2013.

[37] R. Mitchell and I.-R. Chen, "A survey of intrusion detection techniques for cyber-physical systems," *ACM Computing Surveys (CSUR)*, vol. 46, no. 4, p. 55, 2014.

[38] H. Debar, "An introduction to intrusion-detection systems," *Proceedings of Connect*, vol. 2000, 2000.

[39] K. A. Scarfone and P. M. Mell, "Sp 800-94. guide to intrusion detection and prevention systems (idps)," National Institute of Standards & Technology, Report, 2007.

[40] S. Raza, L. Wallgren, and T. Voigt, "Svelte: Real-time intrusion detection in the internet of things," *Ad hoc networks*, vol. 11, no. 8, pp. 2661–2674, 2013.

[41] A. H. Farooqi and F. A. Khan, "Intrusion detection systems for wireless sensor networks: A survey," in *International Conference on Future Generation Communication and Networking*. Springer, 2009, Conference Proceedings, pp. 234–241.

[42] S. O. Amin, M. S. Siddiqui, C. S. Hong, and J. Choe, "A novel coding scheme to implement signature based ids in ip based sensor networks," in *IFIP/IEEE International Symposium on Integrated Network Management-Workshops*. IEEE, 2019, Conference Proceedings, pp. 269–274.

[43] Q. Shen, Z. Zhao, W. Niu, Y. Liu, and H. Tang, "Tolerance-based adaptive online outlier detection for internet of things," in *Proceedings of the 2010 IEEE/ACM Int'l Conference on Green Computing and Communications and Int'l Conference on Cyber, Physical and Social Computing*. IEEE Computer Society, 2010, Conference Proceedings, pp. 560–565.

[44] A. Le, J. Loo, Y. Luo, and A. Lasebae, "Specification-based ids for securing rpl from topology attacks," in *2011 IFIP Wireless Days (WD)*. IEEE, 2011, Conference Proceedings, pp. 1–3.

[45] R. Fu, K. Zheng, D. Zhang, and Y. Yang, "An intrusion detection scheme based on anomaly mining in internet of things," 2011.

[46] C. Liu, J. Yang, R. Chen, Y. Zhang, and J. Zeng, "Research on immunity-based intrusion detection technology for the internet of things," in *Seventh International Conference on Natural Computation*, vol. 1. IEEE, 2011, Conference Proceedings, pp. 212–216.

[47] S. Misra, P. V. Krishna, H. Agarwal, A. Saxena, and M. S. Obaidat, "A learning automata based solution for preventing distributed denial of service in internet of things," in *International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing*. IEEE, 2011, Conference Proceedings, pp. 114–122.

[48] R. Chen, C. M. Liu, and C. Chen, "An artificial immune-based distributed intrusion detection model for the internet of things," in *Advanced materials research*, vol. 366. Trans Tech Publ, 2012, Conference Proceedings, pp. 165–168.

[49] A. Le, J. Loo, A. Lasebae, M. Aiash, and Y. Luo, "6lowpan: a study on qos security threats and countermeasures using intrusion detection system approach," *International Journal of Communication Systems*, vol. 25, no. 9, pp. 1189–1212, 2012.

[50] N. Ruan and Y. Hori, "Dos attack-tolerant tesla-based broadcast authentication protocol in internet of things," in *International Conference on Selected Topics in Mobile and Wireless Networking*. IEEE, 2012, Conference Proceedings, pp. 60–65.

[51] P. Kasinathan, C. Pastrone, M. A. Spirito, and M. Vinkovits, "Denial-of-service detection in 6lowpan based internet of things," in *IEEE 9th international conference on wireless and mobile computing, networking and communications (WiMob)*. IEEE, 2013, Conference Proceedings, pp. 600–607.

[52] P. Kasinathan, G. Costamagna, H. Khaleel, C. Pastrone, and M. A. Spirito, "An ids framework for internet of things empowered by 6lowpan," in *Proceedings of the 2013 ACM SIGSAC conference on Computer and communications security*. ACM, 2013, Conference Proceedings, pp. 1337–1340.

[53] L. Wallgren, S. Raza, and T. Voigt, "Routing attacks and countermeasures in the rpl-based internet of things," *International Journal of Distributed Sensor Networks*, vol. 9, no. 8, p. 794326, 2013.

[54] C. Jun and C. Chi, "Design of complex event-processing ids in internet of things," in *Sixth International Conference on Measuring Technology and Mechatronics Automation*. IEEE, 2014, Conference Proceedings, pp. 226–229.

[55] J.-Y. Lee, W.-C. Lin, and Y.-H. Huang, "A lightweight authentication protocol for internet of things," in *2014 International Symposium on Next-Generation Electronics (ISNE)*. IEEE, 2014, Conference Proceedings, pp. 1–2.

[56] K. Zhang, X. Liang, R. Lu, and X. Shen, "Sybil attacks and their defenses in the internet of things," *IEEE Internet of Things Journal*, vol. 1, no. 5, pp. 372–383, 2014.

[57] P. Pongle and G. Chavan, "Real time intrusion and wormhole attack detection in internet of things," *International Journal of Computer Applications*, vol. 121, no. 9, 2015.

[58] C. Cervantes, D. Poplade, M. Nogueira, and A. Santos, "Detection of sinkhole attacks for supporting secure routing on 6lowpan for internet of things," in *IFIP/IEEE International Symposium on Integrated Network Management (IM)*. IEEE, 2015, Conference Proceedings, pp. 606–611.

[59] R. An, H. Feng, Q. Liu, and L. Li, "Three elliptic curve cryptography-based rfid authentication protocols for internet of things," in *International Conference on Broadband and Wireless Computing, Communication and Applications*. Springer, 2016, Conference Proceedings, pp. 857–878.

[60] H. Sun, X. Wang, R. Buyya, and J. Su, "Cloudeyes: Cloud-based malware detection with reversible sketch for resource-constrained internet of things (iot) devices," *Software: Practice and Experience*, vol. 47, no. 3, pp. 421–441, 2017.

[61] B. Arrington, L. Barnett, R. Rufus, and A. Esterline, "Behavioral modeling intrusion detection system (bmids) using internet of things (iot) behavior-based anomaly detection via immunity-inspired algorithms," in *2016 25th International Conference on Computer Communication and Networks (ICCCN)*. IEEE, 2016, Conference Proceedings, pp. 1–6.

[62] E. Hodo, X. Bellekens, A. Hamilton, P.-L. Dubouilh, E. Iorkyase, C. Tachtatzis, and R. Atkinson, "Threat analysis of iot networks using artificial neural network intrusion detection system," in *2016 International Symposium on Networks, Computers and Communications (ISNCC)*. IEEE, 2016, Conference Proceedings, pp. 1–6.

[63] S. Chen, M. Peng, H. Xiong, and X. Yu, "Svm intrusion detection model based on compressed sampling," *Journal of Electrical and Computer Engineering*, vol. 2016, p. 12, 2016.

[64] A. Le, J. Loo, K. Chai, and M. Aiash, "A specification-based ids for detecting attacks on rpl-based network topology," *Information*, vol. 7, no. 2, p. 25, 2016.

[65] R. J. Colville and G. Spafford, "Configuration management for virtual and cloud infrastructures," *Gartner2010*, 2010.

[66] O. Networking, "Open networking foundation," 2019.

[67] S. Brief, "Sdn security considerations in the data center," 2013.

[68] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "Openflow: enabling innovation in campus networks," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 2, pp. 69–74, 2008.

[69] R. Enns, M. Bjorklund, and J. Schoenwaelder, "Network configuration protocol (netconf)," *Network*, 2011.

[70] A. Doria, J. H. Salim, R. Haas, H. M. Khosravi, W. Wang, L. Dong, R. Gopal, and J. M. Halpern, "Forwarding and control element separation (forces) protocol specification," *RFC*, vol. 5810, pp. 1–124, 2010.

[71] H. Zhong, Y. Fang, and J. Cui, "Lbbsrt: An efficient sdn load balancing scheme based on server response time," *Future Generation Computer Systems*, vol. 68, pp. 183–190, 2017.

[72] N. L. Van Adrichem, C. Doerr, and F. A. Kuipers, "Opennetmon: Network monitoring in openflow software-defined networks," in *2014 IEEE Network Operations and Management Symposium (NOMS)*. IEEE, 2014, Conference Proceedings, pp. 1–8.

[73] V. Mann, A. Vishnoi, K. Kannan, and S. Kalyanaraman, "Crossroads: Seamless vm mobility across data centers through software defined networking," in *2012 IEEE Network Operations and Management Symposium*. IEEE, 2012, Conference Proceedings, pp. 88–96.

[74] A. Tootoonchian, M. Ghobadi, and Y. Ganjali, "Opentm: traffic matrix estimator for openflow networks," in *International Conference on Passive and Active Network Measurement*. Springer, 2010, Conference Proceedings, pp. 201–210.

[75] Z. Latif, K. Sharif, F. Li, M. M. Karim, and Y. Wang, "A comprehensive survey of interface protocols for software defined networks," *arXiv preprint arXiv:1902.07913*, 2019.

[76] J. Brownlee, "Supervised and unsupervised machine learning algorithms," *Machine Learning Mastery*, vol. 16, no. 03, 2016.

[77] E. Hodo, X. Bellekens, A. Hamilton, C. Tachtatzis, and R. Atkinson, "Shallow and deep networks intrusion detection system: A taxonomy and survey," *arXiv preprint arXiv:1701.02145*, 2017.

[78] A. A. Aburomman and M. B. I. Reaz, "Survey of learning methods in intrusion detection systems," in *2016 International Conference on Advances in Electrical, Electronic and Systems Engineering (ICAEES)*. IEEE, 2016, Conference Proceedings, pp. 362–365.

[79] A. Javaid, Q. Niyaz, W. Sun, and M. Alam, "A deep learning approach for network intrusion detection system," in *Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS)*. ICST (Institute for Computer Sciences, Social-Informatics and ..., 2016, Conference Proceedings, pp. 21–26.

[80] S. Zanero and S. M. Savaresi, "Unsupervised learning techniques for an intrusion detection system," in *Proceedings of the 2004 ACM symposium on Applied computing*. ACM, 2004, Conference Proceedings, pp. 412–419.

[81] I. Syarif, A. Prugel-Bennett, and G. Wills, "Unsupervised clustering approach for network anomaly detection," in *International conference on networked digital technologies*. Springer, 2012, Conference Proceedings, pp. 135–145.

[82] C.-F. Tsai, Y.-F. Hsu, C.-Y. Lin, and W.-Y. Lin, "Intrusion detection by machine learning: A review," *expert systems with applications*, vol. 36, no. 10, pp. 11 994–12 000, 2009.

[83] J. Haweliya and B. Nigam, "Network intrusion detection using semi supervised support vector machine," *International Journal of Computer Applications*, vol. 85, no. 9, 2014.

[84] K. P. Bennett and A. Demiriz, "Semi-supervised support vector machines," in *Advances in Neural Information processing systems*, 1999, Conference Proceedings, pp. 368–374.

[85] "Deep learning stand to benefit to data analytics and hpc expertise," 2017. [Online]. Available: https://www.cio.com/article/3180184/deep-learning-stands-to-benefit-from-data-analytics-and-high-performance-computing- hpc-expertise.html

[86] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[87] "Convolutional neural networks," 2014. [Online]. Available: http://eric-yuan.me/cnn/

[88] L. Deng and D. Yu, "Deep learning: methods and applications," *Foundations and Trends® in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[89] M. Z. Alom, V. Bontupalli, and T. M. Taha, "Intrusion detection using deep belief networks," in *2015 National Aerospace and Electronics Conference (NAECON)*. IEEE, 2015, Conference Proceedings, pp. 339–344.

[90] A. Vyas, "Deep learning in natural language processing," 2017.

[91] E. Lear, D. Romascanu, and R. Droms, "Manufacturer usage description specification," *IETF*, 2019.

[92] M. Jethanandani, L. Huang, S. Agarwal, and D. Blair, "Network access control list (acl) yang data model," *IETF Draft*, 2018.

[93] K. Giotis, C. Argyropoulos, G. Androulidakis, D. Kalogeras, and V. Maglaris, "Combining openflow and sflow for an effective and scalable anomaly detection and mitigation mechanism on sdn environments," *Computer Networks*, vol. 62, pp. 122–136, 2014.

[94] C. YuHunag, T. MinChi, C. YaoTing, C. YuChieh, and C. YanRen, "A novel design for future on-demand service and security," in *2010 IEEE 12th International Conference on Communication Technology*. IEEE, 2010, Conference Proceedings, pp. 385–388.

[95] S. A. Mehdi, J. Khalid, and S. A. Khayam, "Revisiting traffic anomaly detection using software defined networking," in *International workshop on recent advances in intrusion detection*. Springer, 2011, Conference Proceedings, pp. 161–180.

[96] K. Giotis, G. Androulidakis, and V. Maglaris, "Leveraging sdn for efficient anomaly detection and mitigation on legacy networks," in *2014 Third European Workshop on Software Defined Networks*. IEEE, 2014, Conference Proceedings, pp. 85–90.

[97] R. Wang, Z. Jia, and L. Ju, "An entropy-based distributed ddos detection mechanism in software-defined networking," in *2015 IEEE Trustcom/BigDataSE/ISPA*, vol. 1. IEEE, 2015, Conference Proceedings, pp. 310–317.

[98] Y. Wang, Y. Zhang, V. K. Singh, C. Lumezanu, and G. Jiang, "Netfuse: Short-circuiting traffic surges in the cloud," in *ICC*. Citeseer, 2013, Conference Proceedings, pp. 3514–3518.

[99] J. Boite, P.-A. Nardin, F. Rebecchi, M. Bouet, and V. Conan, "Statesec: Stateful monitoring for ddos protection in software defined networks," in *2017 IEEE Conference on Network Softwarization (NetSoft)*. IEEE, 2017, Conference Proceedings, pp. 1–9.

[100] S. McGillicuddy, "Radware adds open source ddos protection to opendaylight project," Technical report, Radware, Report, 2013.

[101] H. Peng, Z. Sun, X. Zhao, S. Tan, and Z. Sun, "A detection method for anomaly flow in software defined network," *IEEE Access*, vol. 6, pp. 27 809–27 817, 2018.

[102] R. Braga, E. de Souza Mota, and A. Passito, "Lightweight ddos flooding attack detection using nox/openflow," in *LCN*, vol. 10, 2010, Conference Proceedings, pp. 408–415.

[103] J. Xie, F. R. Yu, T. Huang, R. Xie, J. Liu, C. Wang, and Y. Liu, "A survey of machine learning techniques applied to software defined networking (sdn): Research issues and challenges," *IEEE Communications Surveys and Tutorials*, vol. 21, no. 1, pp. 393–430, 2018.

[104] J. Ashraf and S. Latif, "Handling intrusion and ddos attacks in software defined networks using machine learning techniques," in *2014 National Software Engineering Conference*. IEEE, 2014, Conference Proceedings, pp. 55–60.

[105] M. Glick and H. Rastegarfar, "Scheduling and control in hybrid data centers," in *IEEE Photonics Society Summer Topical Meeting Series (SUM)*. IEEE, 2017, Conference Proceedings, pp. 115–116.

[106] C. Song, Y. Park, K. Golani, Y. Kim, K. Bhatt, and K. Goswami, "Machine-learning based threat-aware system in software defined networks," in *26th international conference on computer communication and networks (ICCCN)*. IEEE, 2017, Conference Proceedings, pp. 1–9.

[107] Q. Niyaz, W. Sun, and A. Y. Javaid, "A deep learning based ddos detection system in software-defined networking (sdn)," *arXiv preprint arXiv:1611.07400*, 2016.

[108] T. A. Tang, L. Mhamdi, D. McLernon, S. A. R. Zaidi, and M. Ghogho, "Deep learning approach for network intrusion detection in software defined networking," in *2016 International Conference on Wireless Networks and Mobile Communications (WINCOM)*. IEEE, 2016, Conference Proceedings, pp. 258–263.

[109] S. Lee, J. Kim, S. Shin, P. Porras, and V. Yegneswaran, "Athena: A framework for scalable anomaly detection in software-defined networks," in *2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 2017, Conference Proceedings, pp. 249–260.

[110] M. Ranganathan, "Soft mud: Implementing manufacturer usage descriptions on openflow sdn switches," in *International Conference on Networks (ICN)*, 2019, Conference Proceedings.

[111] A. Hamza, H. H. Gharakheili, T. A. Benson, and V. Sivaraman, "Detecting volumetric attacks on lot devices via sdn-based monitoring of mud activity," in *Proceedings of the 2019 ACM Symposium on SDN Research*, 2019, pp. 36–48.

[112] A. Hamza, H. H. Gharakheili, and V. Sivaraman, "Combining mud policies with sdn for iot intrusion detection," in *Proceedings of the 2018 Workshop on IoT Security and Privacy*, 2018, pp. 1–7.

[113] A. Hamza, D. Ranathunga, H. H. Gharakheili, M. Roughan, and V. Sivaraman, "Clear as mud: Generating, validating and applying iot behavioral profiles," in *Proceedings of the 2018 Workshop on IoT Security and Privacy*, 2018, pp. 8–14.

[114] J. Li, S. Berg, M. Zhang, P. Reiher, and T. Wei, "Drawbridge: Software-defined ddos-resistant traffic engineering," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 4, pp. 591–592, 2014.

[115] M. Yu, Y. Zhang, J. Mirkovic, and A. Alwabel, "SENSS: Software defined security service," in *Presented as part of the Open Networking Summit 2014 (ONS 2014)*, 2014, Conference Proceedings.

[116] S. K. Fayaz, Y. Tobioka, V. Sekar, and M. Bailey, "Bohatei: Flexible and elastic ddos defense," in *24th USENIX Security Symposium (USENIX Security 15)*, 2015, Conference Proceedings, pp. 817–832.

[117] R. Sahay, G. Blanc, Z. Zhang, and H. Debar, in *Towards autonomic DDoS mitigation using software defined networking*, 2015, Conference Proceedings.

[118] ——, "Aroma: An sdn based autonomic ddos mitigation framework," *Computers and Security*, vol. 70, pp. 482–499, 2017.

[119] M. Wichtlhuber, R. Reinecke, and D. Hausheer, "An sdn-based cdn/isp collaboration architecture for managing high-volume flows," *IEEE Transactions on Network and Service Management*, vol. 12, no. 1, pp. 48–60, 2015.

[120] L. Von Ahn, M. Blum, N. J. Hopper, and J. Langford, "Captcha: Using hard ai problems for security," in *International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2003, Conference Proceedings, pp. 294–311.

[121] D. Li, L. Deng, M. Lee, and H. Wang, "Iot data feature extraction and intrusion detection system for smart cities based on deep migration learning," *International Journal of Information Management*, 2019.

[122] D. G. Gonçalves, F. L. de Caldas Filho, L. M. E. Martins, G. d. O. Kfouri, B. V. Dutra, R. d. O. Albuquerque, and R. T. de Sousa, "Ips architecture for iot networks overlapped in sdn," in *2019 Workshop on Communication Networks and Power Systems (WCNPS)*. IEEE, 2019, Conference Proceedings, pp. 1–6.

[123] S. Singh, A. Atrey, M. L. Sichitiu, and Y. Viniotis, "Clearer than mud: Extending manufacturer usage description (mud) for securing iot systems," in *International Conference on Internet of Things*. Springer, 2019, Conference Proceedings, pp. 43–57.

[124] A. Kumar and T. J. Lim, "Edima: Early detection of iot malware network activity using machine learning techniques," *arXiv preprint arXiv:1906.09715*, 2019.

[125] P. Amangele, M. J. Reed, M. Al-Naday, N. Thomos, and M. Nowak, "Hierarchical machine learning for iot anomaly detection in sdn," 2019.

[126] A. Wani and S. Revathi, *Analyzing Threats of IoT Networks Using SDN Based Intrusion Detection System (SDIoT-IDS)*. Singapore: Springer Singapore, 2018, vol. 828, pp. 536–542.

[127] L. Chang, "A proactive approach to detect iot based flooding attacks by using software defined networks and manufacturer usage descriptions," Thesis, Arizona State University, 2018.

[128] C. Wu, Y. Liu, F. Wu, F. Liu, H. Lu, W. Fan, and B. Tang, "A hybrid intrusion detection system for iot applications with constrained resources," *International Journal of Digital Crime and Forensics (IJDCF)*, vol. 12, no. 1, pp. 109–130, 2020.

[129] S. Venkatraman and B. Surendiran, "Adaptive hybrid intrusion detection system for crowd sourced multimedia internet of things systems," *Multimedia Tools and Applications*, vol. 79, no. 5, pp. 3993–4010, 2020.

[130] Y. N. Soe, Y. Feng, P. I. Santosa, R. Hartanto, and K. Sakurai, "Towards a lightweight detection system for cyber attacks in the iot environment using corresponding features," *Electronics*, vol. 9, no. 1, p. 144, 2020.

[131] G. D. Putra, V. Dedeoglu, S. S. Kanhere, and R. Jurdak, "Towards scalable and trustworthy decentralized collaborative intrusion detection system for iot," *arXiv preprint arXiv:2002.07512*, 2020.

[132] Y. Li, Y. Xu, Z. Liu, H. Hou, Y. Zheng, Y. Xin, Y. Zhao, and L. Cui, "Robust detection for network intrusion of industrial iot based on multi-cnn fusion," *Measurement*, vol. 154, p. 107450, 2020.

[133] M. A. Ferrag, L. Maglaras, A. Ahmim, M. Derdour, and H. Janicke, "Rdtids: Rules and decision tree-based intrusion detection system for internet-of-things networks," *Future Internet*, vol. 12, no. 3, p. 44, 2020.

[134] T. D. Nguyen, S. Marchal, M. Miettinen, H. Fereidooni, N. Asokan, and A.-R. Sadeghi, "DÏot: A federated self-learning anomaly detection system for iot," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, Conference Proceedings, pp. 756–767.

[135] C. Cervantes, D. Poplade, M. Nogueira, and A. Santos, "Detection of sinkhole attacks for supporting secure routing on 6lowpan for internet of things," in *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*. IEEE, 2015, Conference Proceedings, pp. 606–611.

[136] M. Eskandari, Z. H. Janjua, M. Vecchio, and F. Antonelli, "Passban ids: An intelligent anomaly based intrusion detection system for iot edge devices," *IEEE Internet of Things Journal*, 2020.

[137] P. Bhale, S. Dey, S. Biswas, and S. Nandi, "Energy efficient approach to detect sinkhole attack using roving ids in 6lowpan network," in *International Conference on Innovations for Community Services*. Springer, 2020, Conference Proceedings, pp. 187–207.

[138] M. J. Babu and A. R. Reddy, "Sh-ids: Specification heuristics based intrusion detection system for iot networks," *Wireless Personal Communications*, pp. 1–23, 2020.

[139] B. Al-Duwairi, W. Al-Kahla, M. A. AlRefai, Y. Abdelqader, A. Rawash, and R. Fahmawi, "Siem-based detection and mitigation of iot-botnet ddos attacks," *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 10, 2020.

[140] A. Mudgerikar, P. Sharma, and E. Bertino, "E-spion: A system-level intrusion detection system for iot devices," in *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, 2019, Conference Proceedings, pp. 493–500.

[141] V. Kumar, A. K. Das, and D. Sinha, "Uids: a unified intrusion detection system for iot environment," *Evolutionary Intelligence*, pp. 1–13, 2019.

[142] D. Breitenbacher, I. Homoliak, Y. L. Aung, N. O. Tippenhauer, and Y. Elovici, "Hades-iot: A practical host-based anomaly detection system for iot devices," in *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, 2019, Conference Proceedings, pp. 479–484.

[143] J. Straub, "An interdiction detection and prevention system (idps) for anti-autonomy attack repulsion," in *2019 IEEE Aerospace Conference*. IEEE, 2019, Conference Proceedings, pp. 1–8.

[144] P. R. Chandre, P. N. Mahalle, and G. R. Shinde, *Deep Learning and Machine Learning Techniques for Intrusion Detection and Prevention in Wireless Sensor Networks: Comparative Study and Performance Analysis*. Springer, 2020, pp. 95–120.

[145] S. Alam, M. Shuaib, and A. Samad, "A collaborative study of intrusion detection and prevention techniques in cloud computing," in *International Conference on Innovative Computing and Communications*. Springer, 2019, Conference Proceedings, pp. 231–240.

[146] K. Scarfone and P. Mell, *Intrusion detection and prevention systems*. Springer, 2010, pp. 177–192.

[147] B. JOSEFSSON, *Securing your Industrial IoT ecosystem against cyber threats*, 2019. [Online]. Available: https://www.ericsson.com/en/blog/2019/4/securing-your-industrial-IoT-network-against-cyber-threats

[148] S. Arunkumar, S. Pipes, C. Makaya, E. Bertino, E. Karafili, E. Lupu, and C. Williams, "Next generation firewalls for dynamic coalitions," in *2017 IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computed, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, 2017, Conference Proceedings, pp. 1–6.

[149] F. Hussain, R. Hussain, S. A. Hassan, and E. Hossain, "Machine learning in iot security: Current solutions and future challenges," *arXiv preprint arXiv:1904.05735*, 2019.

[150] R. Sahay, W. Z. Meng, and C. D. Jensen, "The application of software defined networking on securing computer networks: A survey," *Journal of Network and Computer Applications*, vol. 131, pp. 89–108, 2019.

[151] A. Bahga and V. K. Madisetti, "Blockchain platform for industrial internet of things," *Journal of Software Engineering and Applications*, vol. 9, no. 10, p. 533, 2016.

[152] K. Christidis and M. Devetsikiotis, "Blockchains and smart contracts for the internet of things," *Ieee Access*, vol. 4, pp. 2292–2303, 2016.

[153] C. Iwendi, Z. Jalil, A. R. Javed, T. Reddy, R. Kaluri, G. Srivastava, and O. Jo, "Keysplitwatermark: Zero watermarking algorithm for software protection against cyber-attacks," *IEEE Access*, vol. 8, pp. 72 650–72 660, 2020.

Multi Modal RGB D Action Recognition with CNN LSTM Ensemble Deep Network

D. Srihari[1] , P. V. V. Kishore[2]
Department of ECE, Koneru Lakshmaiah Education Foundation,
Guntur (DT), Andhra Pradesh, India

Abstract—**Human action recognition has transformed from a video processing problem into multi modal machine learning problem. The objective of this work is to perform multi modal human action recognition on an ensemble hybrid network of CNN and LSTM layers. The proposed CNN - LSTM ensemble network is a 2 - stream framework with one ensemble stream learning RGB sequences and the other depth. This proposed framework can learn both temporal and spatial dynamics in both RGB and depth modal action data. The hybrid network is found to be receptive towards both spatial and temporal fields because of the hierarchical structure of CNNs and LSTMs. Finally, to test our proposed model, we used our own BVCAction3D and three RGB D benchmark action datasets. The experiments were conducted on all the datasets using the proposed framework and was found to be effective when compared to similar deep learning architectures.**

***Keywords*—Human action recogniiton; RGB D video data; convolutional neural networks; long short-term memory**

I. Introduction

Human action recognition is basically considered as a computer vision problem where a set of video processing algorithms were proposed to extract features that became input to a classification algorithm. However, these video processing algorithms depended heavily on the orientations of pixels in the video frames which affected the performance of the classifier as whole. Despite their instabilities in generalizing the classifiers performance the human action recognition are applied in surveillance networks, industrial automation, medical and sports analysis to name a few. In contrast to RGB video sensors, we now have low cost multi modal sensors such as Microsoft Kinect, that can enhance RGB sequences with depth and skeletal information. On the other hand the progress of deep learning algorithms like Convolutional Neural Networks (CNNs) and Recurrent models (RNNs) has been instrumental in enhancing the performance of multi modal action recognition systems.

In the recent years deep learning architectures have been shown to learn and complement the unique features in RGB, depth and skeletal data for performance improvements in action recognition tasks [1], [2]. Specifically, the work in [3] shows the effectiveness of using auxiliary datasets in the form of skeleton and depth has enhanced the accuracy of action recognition system using RGB videos. Multiple Kernel based learning framework was applied effectively on RGB D action data for extracting multi modal features and further fusing them, which improved their accuracy positively [4]. Further, a few of these models explored the sparse modelling of dense RGB and depth features that were translated into weighted bag of words (BOW) [5] representation for classification. Most of the works experimented with full action sequences ignoring the temporal information accompanying the action.

Initially, the idea was to extract motion information from RGB and depth sequences using optical flow, Kalman tracking and sometimes packing motion into a single image called as motion history images (MHI) [6]. Even though these methods offered an improvement in performance of the classifiers, they showed difficulty in learning spatio temporal features for generalizing an action. The fundamental difficulty in multi modal sequences is the formation of a multi-dimensional tensor indexing modalities, their spatial and temporal knowledge in one field. Subsequently, learning and temporal pooling operations on this multi-dimensional tensor is a challenging task. Moreover, time varying modalities will always induce constraints due to variable length effects. Despite the above gaps in data acquisition and processing, the time varying multi modal features can enhance the performance of the learning algorithms. However, the question posed at this instance is how to teach a classifier the spatio temporal modalities for RGB D action recognition tasks.

Previously, we approached the above problem by dividing multiple modalities to fixed length action sequences which are then arranged as a multi layered multi modal tensor. These multi-dimensional tensors are processed through deep convolutional neural networks (CNN) for learning spatial representations thereby completely ignoring the temporal structures [7].

In this paper, we propose to develop a hybrid recurrent CNN based deep learning framework for multi modal action recognition from RGB and depth data. Our proposed CNN LSTM network has been an inception of recurrent CNNs for action recognition in [8]. However, it is different from models proposed previously on multi modal action sequences [9], [10], [11], [12] in two aspects. One, the multi modelled data used in our work is RGB and depth sequences and two, our proposed CNN-LSTM Action Network (CLANet) is an ensemble of streams of layers.

The CLANet extracts the spatial features from RGB and depth sequences using CNN and infuses the extracted features into the LSTM network which is bidirectional in structure. The LSTM streams learn the temporal patterns in both the RGB and depth sequences during training. The last layer of the CLANet is a dense layer with SoftMax activation the outputs of which are score fused to decide on the input class. We opted for RGB and depth sequences for this experimentation and no skeletal action inputs due to the data dimensionality representation between them. Skeletal data has a higher dimensionality over the RGB and depth which share a common representation.

In order to validate our proposed framework, we have our own BVC3DA RGB D action dataset with 40 actions from 10 different actors with 10 repetitions per action. However, we evaluated the proposed framework on benchmark RGB D datasets, NTU RGB D, MSRAction and UTKINECT to test the learning strategies of the proposed CNN LSTM network.

The rest of the paper is organized as follows. The second section presents the previous works on RGB D multi modal action recognition with an insight into gaps and the achieved breakthroughs. Section three discusses the methods applied in this study to achieve higher performances across datasets for multimodal action recognition. Results are presented in section four and conclusions drawn on the obtained results in section five.

II. Background

Multi modal or RGB D based action recognition models has been studied extensively which led to the development of the proposed CLANet. The previous methods have shown to have used data with RGB frames, depth and skeletal information for recognition of human actions across multiple identification platforms such as machine learning [1] to deep learning [13]. The machine learning models apply segmentation and feature extraction algorithms on RGB or depth or both frames for extracting meaningful representations of actions [14]. On the other hand, deep learning models extract features and segments based on the training algorithms on the RGB D data sequences [9]. The most formidable of these deep learning models are grouped into spatial and temporal domains. In spatial domain the models extract features with respect to the pixel location in image space using models such as Convolutional Neural Networks (CNNs) [2], [7]. For temporal or time series modelling of the RGB D data, Recurrent neural networks (RNNs) and their upgrades such as Long Short-Term Memory (LSTM) nets [15], [16].

However, the spatial and temporal models have their share of advantages and disadvantages. The spatial models cannot effectively learn the time series information which is necessary to represent action sequences that dependent on continuous data variations. Contrastingly, using exclusive time series modelling on video frames will not capture the spatial representations of action movements in image spaces. Hence, a hybrid combination involving both spatial and temporal models in found to be necessary to represent actions in video sequences for recognition [17], [18]. The early models applied optical flow to extract the temporal features on RGB video frames which are further fused with the spatial features during the training of CNNs. A few state-of-the-art models used multiple streams of independent CNNs with inputs from RGB and optical flow based RGB giving satisfactory results [19], [20]. One stream of CNN used RGB spatial features and the other uses motion information during training the networks simultaneously. All these networks are accompanied with feature fusion layer before or after the dense layers for decision making on the inputted action sequence. However, these models require additional computation time in the form of motion vectors which makes them computationally inefficient due to data alignment problems. Moreover, few also tried 4 streams by adding motion information from depth sequences producing better recognition accuracies than the previous 2 stream model [7].

Similar to the above models, properties of the RGB and depth modalities have produced efficient action recognition algorithms such as depth rank pooling with CNNs [21], scene flow based RGB D channels on CNN [22] and sequence based methods with RNNs [23]. However, the most successful are models that combine the advantages of both spatial and temporal networks. These models are named as spatio temporal recurrent convolutional neural networks (rCNNs) [24]. These models operate in twofold: one, the primary network extracts the spatial features using CNNs and the secondary network encodes that spatial features into temporal data using recurrent models. The most frequently applied recurrent model was Long short-term memory (LSTM) for representing temporal information in the action video sequences due to their ability handle long term dependencies by avoiding gradient vanishing problems [25]. Consequently, it was found that the operating the feature pooling model with LSTM can influence the temporal learning capabilities of the hybrid CNN LSTM architectures. Through feature sharing mechanism between the two networks, they were able to produce higher level representations of actions in a video sequence [26]. Moreover, bidirectional LSTM based methods have shown to handle multiple length video sequences when compared to RNNs. Therefore, the hybrid combination of CNN and LSTMs is the most widely applied model for human action recognition because of their abilities to decode spatial and temporal information simultaneously [27].

Literature is filled with CNN LSTM models for action recognition using skeletal actions as inputs [28], [29], [30]. These models use 3D skeletal joints as time series data along with RGB video frames for training and testing. However, depth-based models were rarely used along with these hybrid models [31]. In this work, we try to learn through a hybrid model which uses both RGB and depth data to draw inferences on the input action sequences. Both CNNs and LSTMs allow end to end trainable models that eliminates the need for tracking variations through time series data. The advantages of using RGB inputs along with depth instead of skeletal data are threefold. First, the depth features are more profound in assisting the spatial information in RGB data when compared to skeletal data. Second, the depth data is analogous to RGB data, which allows for complex processing mechanisms in transforming the skeletal data to image data. Finally, the skeletal data at times is found to be noisy with missing joints or overlapping joints making it difficult to process.

Eventually, in this work we describe a hybrid framework by combining LSTMs with CNNs for action recognition called as CLANet to construct an end-to-end trainable architecture that has capabilities in handling visual action recognition and sequence prediction tasks.

III. Methodology

This section provides a detailed description of the proposed CLANet hybrid CNN - LSTM architecture for action recognition. First, we design a deep CNN model to extract RGB and depth features of multiple frames to generate spatial features in the considered RGB and depth modals respectively. Then we will build an end-to-end pipeline architecture by combining multi modal CNNs with bi-directional LSTMs, followed by a

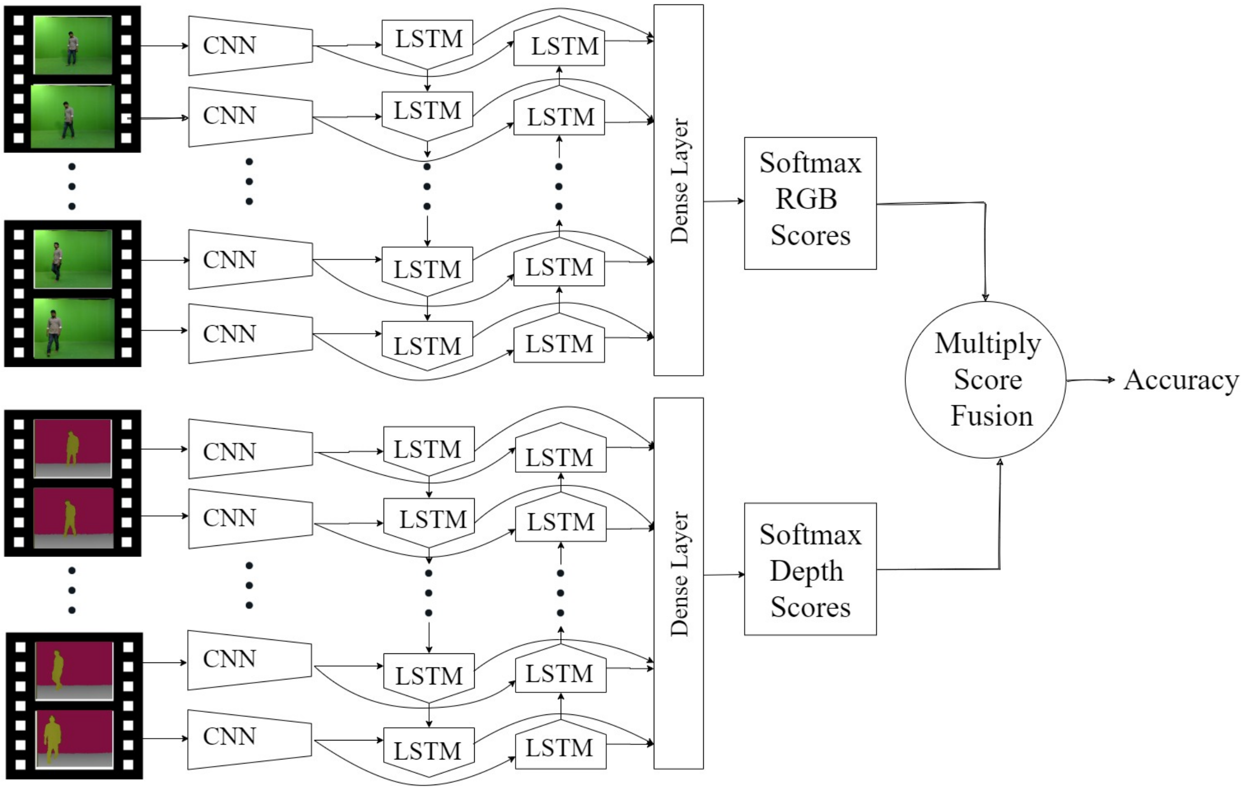


Fig. 1. Proposed CLANet Architecture for RGB D action recognition.

multiply score fusion to estimate the actions. The proposed architecture is shown in Fig. 1.

A. The Spatial CNN Network

This subsection describes in detail the architecture for extracting spatial information from RGB and depth video frames. To accomplish this, we employed convolutional neural networks in multiple streams that take input as RGB and depth video frames. Based on the GPU memory, we found that the maximum number of streams that can be applied in a batch is 16. Hence, the first hyperparameter selected was batch size which is set to 16. Hence each ensemble of CNNs will feed into 16 frames of RGB and depth frames. Lets name the two ensembles are CRGBe and CDe. The CRGBe and CDe are multi stream ensembles of CNNs for RGB inputs and depth inputs, respectively. Fig. 2 shows the CNN architecture developed for extracting spatial features from RGB video frames. Consequently, we have a similar network, CDe for processing depth frames.

Given an RGB action video frame $V_{rgb}\left(v_r, v_g, v_b\right)$ with a pixel position of (x, y), the output of 2D convolutional kernels are feature maps. Eventually, the j^{th} feature map from the i^{th} convolutional layer is extracted using the expression

$$F_{ij}\left(x,y\right) = f\left(\sum_{p}\sum_{n=0}^{N-1}\sum_{m=0}^{M-1}\left(W_{ijp}^{nm} * V_{(i-1)p}\left(x+n, y+m\right)\right)+b_{ij}\right) \quad (1)$$

Where, $N \times M$ is the size of the video frame $V$and f is the activation function. W_{ijp}^{nm} is the weight vector at position (n, m) associated with p^{th} feature map in the $(i-1)^{th}$ layer of the CNN network. The parameter b_{ij} is the bias associated with each of the neurons. Eq'n(1) depicts the convolutional operation between the video frames and the weight matrix, which is updated sequentially during training of the network. There are 16 streams in CRGBe ensemble network to extract spatial features of 16 consecutive frames per action video. To maintain uniformity, we divided each action class video into 128 frames. That is there will 8 batches of RGB video frames per class for training on the CRGBe network. Subsequently, the depth network, CDe will also have the same configuration as CRGBe. The CDe extracts the depth spatial features from the depth sequences of actions.

As shown in Fig. 2, the architecture for RGB spatial feature extraction module with RGB input video frames. The CRGBe is an ensemble of 16 streams with a depth of 10 layers each. The 10 layers depth across each stream consists of 6 convolutional plus ReLu layers, 3 max pooling layers and a flatten layer. The filter kernels are selected as 7×7, 5×5 and 3×3 framework. This kernel selection framework has ensured a hierarchical feature extraction model that has ensured maximal spatial preservation of pixels towards the end of the network. Similar functionality is achieved on depth frames using CDe network. The spatial maps from CRGBe and CDe are now used for modelling temporal information in the features by passing them through LSTM module. The LSTM module is presented in the following section.

B. The LSTM Temporal Coding

The extracted spatial features from the two ensemble nets, CRGBe and CDe, are then temporally coded for recognition

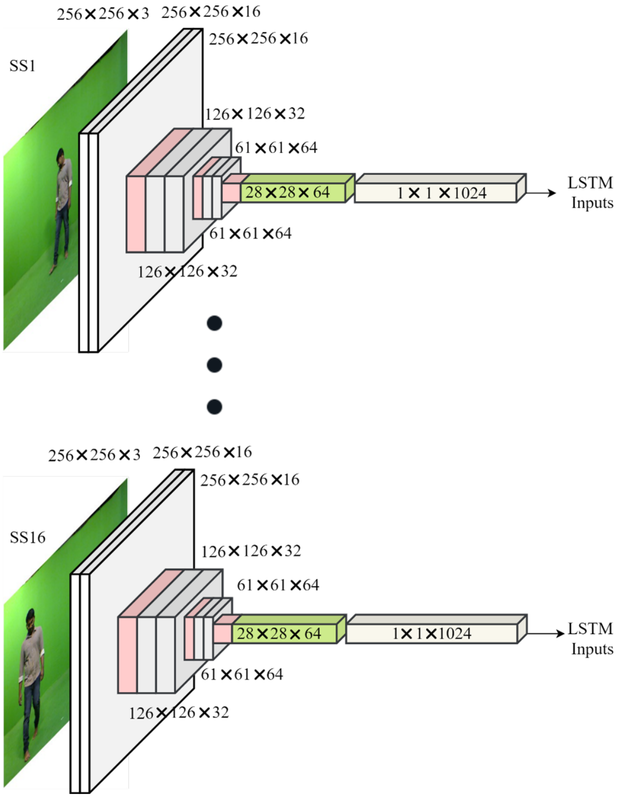


Fig. 2. The CRGBe ensemble for extracting spatial features from RGB action video frames.

of actions at the highest level. LSTM blocks provide temporal dynamics for the extracted spatial features across both the input modalities. Fig. 3 illustrates the single LSTM block used in this work. The following expressions are implemented during the operation of an LSTM block. It consists of an input unit I_t, forget gate F_t, output gate O_t, momentum factor G and the LSTM cell outputs (C_t, h_t).

$$I_t = \sigma\left(\left(x_t + h_{t-1}\right) W_I + b_I\right) \tag{2}$$

$$F_t = \sigma\left(\left(x_t + h_{t-1}\right) W_F + b_F\right) \tag{3}$$

$$O_t = \sigma\left(\left(x_t + h_{t-1}\right) W_O + b_O\right) \tag{4}$$

$$G = \tanh\left(\left(x_t + h_{t-1}\right) W_G + b_G\right) \tag{5}$$

$$C_t = C_{t-1} * F_t + G * I_t \tag{6}$$

$$h_t = O_t * \tanh\left(C_t\right) \tag{7}$$

Where, W and b are weights and bias. x_t are the feature inputs extracted using the spatial CNN network.C_t andh_t are LSTM's cell state at time step t. The sigmoid σ acts as control gates for transfer of inputs to the outputs. The forget gate initiates the progress of inputs to the next LSTM block. Based on the state of forget gate, the LSTM cell either forgets or memorizes the features in a sequence. However, the flow is unidirectional in a single stream LSTM model. In general, a sequence labelling problem such as video-based action recognition we need access to the past and future inputs at a single time step during the training sequence. This is found to be achievable in the past using bidirectional LSTM networks as shown in Fig. 1. This is performed by two LSTM streams with one moving past data forward and the other moving the future data backwards for a specific time step. This biLSTM network is also trained using the same backpropagation through time algorithm. In our work, we performed the backward and forward passes for each action sequence. Subsequently, the hidden states of LSTMs were reset after each action class. Our work uses a bidirectional LSTM architecture from [25]. The following subsection describes the complete multi modal action recognition framework with bidirectional LSTM network on top of CNN networks.

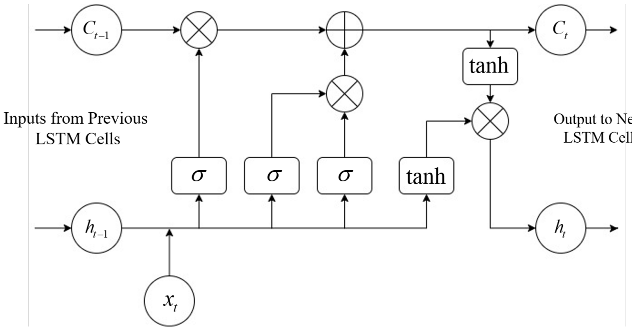


Fig. 3. A Single LSTM Cell Architecture

C. The Hybrid CLANet Training

The hybrid CLANet is designed by stacking bidirectional LSTM cells on top of spatial CNNs to create an end-to-end trainable model. The CNNs are capable of extracting global and highly discriminating spatial features from the RGB and depth video frames. On the other, LSTM capture the local and time representations in the extracted features. Finally, the outputs of LSTM network is passed through a dense layers and a SoftMax layer to compute the probabilistic distribution of the class labels as

$$Y_{class} = SoftMax\left(h_t\right) \tag{8}$$

In the proposed bidirectional LSTM, the hidden states from forward pass and backward pass are combined in the output dense layer. We used 2 dense layers of sizes 1024 each along with a SoftMax to compute the recognition scores. The validation losses are calculated after the first dense layer to update the weights and biases through backpropagation. The validation data is 15% of the total training data and cross entropy loss is used for error calculation. The hyper parameters such as weights and biases are selected randomly with zero mean random gaussian generator. Stochastic gradient descent algorithm is used calculating the losses during training with an initial learning rate of 0.001 across all datasets. However, the learning rate is readjusted, whenever the loss became constant during training. The entire CLANet is end-to-end trainable.

D. The CLANet Testing

The action datasets are divided into 65% training, 15% validation and 20% testing. The outputs of the network give a probability distribution across classes for a particular test input sample. We used multiple machines for training and testing at different frame rates to understand the characteristics of the CLANet in processing multi model spatio temporal data.

Finally, we perform multiple test mechanisms on our own BVRCAction3D dataset and other benchmark datasets such as MSRDailyActivity3D, UTKinect and NTU RGB D.

IV. Results and Analysis

This section presents results of experimentation with analysis of various components that were instrumental in generating the results on various datasets. We start by describing the datasets for training, validation and testing. Next, we initiate the training and testing of the proposed CLANet across different actions in our dataset. Subsequently, we apply benchmark datasets on CLANet for inspecting its rationale against our dataset. Finally, we compare our CLANet with other state-of-the-art multi stream CNN LSTM models for cross data action recognition.

A. Datasets and Performance Measures

The NTU RGB D [32] is the largest dataset with 60 action classes in 80 views recorded with 40 subjects with a total sample size of 56880 videos of skeleton, depth and RGB. We selected 60 action classes with 40 subjects for training and testing the proposed CLANet. The NTU RGB D dataset used in our work has 2400 video samples with 40 subjects in 60 action classes. MSRDailyActivity3D [33] is another standard benchmark dataset using Microsoft Kinect with 16 activity types. It consists of 320 video samples in both RGB and depth modes with actions performed in both sitting and standing positions. The other most widely used RGB D action dataset for benchmarking is UTKinect [34] which has 10 actions from 10 subjects each performing the action twice. It has 10 classes with $10 \times 10 \times 2 = 200$ videos of both RGB and depth data. Inspired from the above benchmark datasets, we collected our own BVRCAction3D action dataset with 40 single human and 10 two human actions using 5 subjects. The complete list of actions is available at [7]. Fig. 4 presents some action sequences in RGB and depth from our BVRCAction3D dataset.

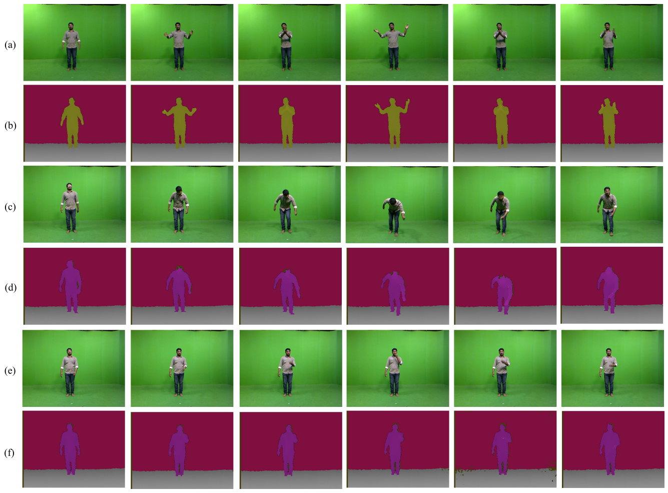


Fig. 4. BVRCAction3D Dataset. Sample RGB and Depth Video Frames of (a)-(b): Clapping, (c)-(d): Mopping the Floor and (e)-(f): Eating.

Our BVRCAction3D dataset consists of $50 \times 5 \times 2 = 500$ video sequences with 50 classes from 5 subjects each performing the action twice. We used Kinect 1.0 for capturing the actions. Each action was recorded for 60 seconds at 30fps. Consequently, each action video has 1800 frames with a resolution of 640×480 for RGB and 320×240 for depth. In order to maintain uniformity across datasets, we resized the frame sizes to 256×256 in both RGB and depth modal videos. Moreover, we found the number of frames in each video clip to have a high degree of similarity among themselves. To increase the redundancy in the action videos, we selected 120 Key frames per action by applying correlation based key frame extractor [35].

The performance of the proposed deep network is measured using two standard parameters: mean Recognition Accuracy (nRA) and mean f1 score (mf1). Apart from the two, we also obtained confusion matrices and region of convergence (ROC) plots across all datasets. In the following subsection, we apply various datasets to our proposed CLANet and evaluate its performance.

B. CLANet Performance

The proposed multi modal CLANet is trained with RGB D action sequences from our BVRCAction3D and other benchmark datasets. The training parameters were kept constant across dataset to understand the implications of data on the network. The hyperparameters of the network were selected as discussed in the previous section. Fig. 5 shows the confusion matrices on the datasets used in this work. The performance of CLANet on our dataset is high when compared to other datasets due to less noisy backgrounds in BVRCAction3D as shown in the Fig. 4. The scores from CLANet are found to be better than our previous work in [7], where we used multi stream CNN with motion information. The reason for higher accuracies is because of the LSTM network which models the time series information in a more accurately. The testing in this case in preformed with 10 test samples only.

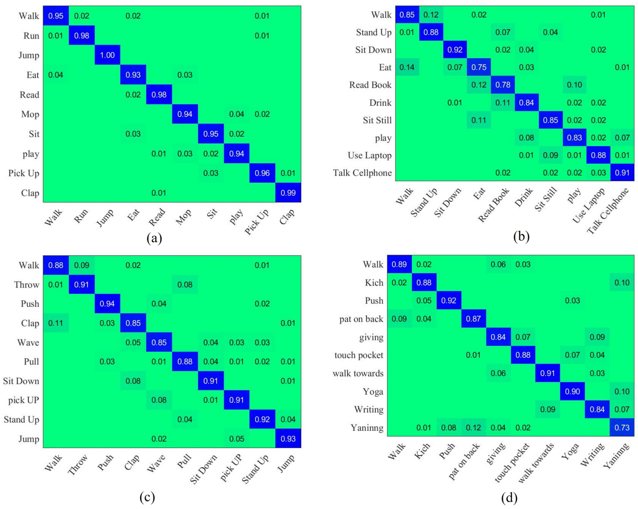


Fig. 5. Confusion Matrices for CLANet on (a) BVRCAction3D, (b) MSRDailyActivity3D, (c) UTKinect and (d) NTU RGB D Datasets for 10 Test Samples.

Eventually, we tested the trained CLANet with the entire testing dataset from each dataset and projected the results in Table I. The results in Table I indicate two performance parameters mRA and mF1 for the proposed CLANet across multiple datasets used in this work. The testing is conducted in cross subject mode, means that the network is shown samples with subjects that are previously unseen by the network during training. The average recognition rate achieved is around 93.32% on our BVRCAction3D dataset, which is found to be better than our previous work in [7].

The above comparison with our work in [7] is important in the context of understanding the need for time series modelling against motion modelling using optical flow. Contrastingly, optical flow-based motion estimation and processing it with regular spatial network has limitations in characterizing the changes across multiple frames. Additionally, the flow-based models fail to capture the long-term dependencies in the action video sequences. Interestingly, hybrid CNN LSTM networks have performed exceptionally well by modelling spatio temporal contents in the action video sequences. Meanwhile, the depth data has come in to assist this process by increasing the performance of the network.

However, it is not possible to generate depth data in real time and hence, we conducted a RGB only test on our proposed CLANet to understand its usability as a real time application. We supplied zero matrices in place of depth data during testing for the depth stream. This test resulted in a mean accuracy of 84.76% and a mean f1 score of 0.862 for BVRCAction3D dataset. The second right half of table I shows the results on all datasets. In spite of depth data absence during testing, the proposed CLANet has performed better on our BVRCAction3D dataset when compared to other benchmark datasets. Consequently, the performance of the network has to be gauged by comparing its performance against state-of-the-art networks as presented in the next subsection.

C. Comparison with Recurrent Hybrid Networks

This subsection gives the comparison of hybrid CNN LSTM networks with RGB and depth inputs as training data. Surprisingly, there are very few works which used both RGB and depth data with hybrids networks for action recognition applications. However, there are a large contingent of networks for skeleton based action recognition using CNN LSTM architectures. Table II presents the comparison of our proposed CLANet with the previously proposed methods for action recognition using the benchmark datasets. We implemented all these networks on the datasets and the mean average recognition is calculated across the training data. The results show that the proposed network outperforms the existing models. All the hyper parameters of the networks were incepted from the proposed CLANet. This is because of the spatio temporal characteristics that are learned effectively by the network in two modalities simultaneously. However, it would be interesting to check the network performance against different action recognition models. Hence, in the next subsection, we compare our method with other state-of-the-art RGB D based action recognition models.

D. Comparison with RGB D Action Recognition Deep Models

The parameters used for training and testing were as described in Section III. In all experiments, the video resolutions were fixed at $256\times256\times3$ for both training and testing for both

TABLE I. Performance of CLANet across datasets and comparison with work from [7]

Dataset	RGB and Depth Testing				RGB only testing			
	mRA	mf1	mRA [7]	mf1 [7]	mRA	Mf1	mRA [7]	Mf1 [7]
BVRCAction3D	93.32	0.965	92.05	0.942	84.76	0.862	72.26	0.687
MSRDailyActivity3D	88.42	0.924	84.86	0.876	74.89	0.784	62.24	0.638
UTKinect	90.25	0.937	87.63	0.894	75.33	0.795	63.85	0.643
NTU RGB D	91.42	0.948	90.27	0.912	78.96	0.812	66.11	0.672

TABLE II. Comparison of average recognition (mRA) of Hybrid recurrent CNN models with RGB and Depth inputs.

Method	Modality	BVRCAction3D	MSRDailyActtivity3D	UTKinect	NTU RGB D
Deep Bilinear CNN [1]	RGB+Depth	82.36	74.8	76.24	79.2
Auxiliary Dataset Model [3]	RGB+Depth+Skeletal maps	92.22	88.12	88.36	89.36
Optical Flow CNN [7]	RGB+Depth	92.05	84.86	87.63	90.27
2 Stream CNN [4]	RGB+Depth	80.23	94.53	95.23	96.32
CLANet Proposed	RGB+Depth	93.32	88.42	90.25	91.42

TABLE III. Comparison of various deep learning models for RGB D action recognition

Modality	Methods	mRA			
		BVRCAction3D	MSRDailyActtivity3D	UTKinect	NTU RGB D
RGB	Two Stream CNN [36]	69.22	68.32	69.03	69.96
	CNN LSTM	73.34	70.82	71.98	73.89
	Spatio Temporal CNN [19]	71.89	69.33	71.12	72.02
	3D CNN LSTM [29]	78.96	74.02	75.64	78.91
Depth	CNN [21]	71.84	68.96	69.98	70.52
	CNN [22]	72.36	70.50	71.22	72.36
	RNN [23]	74.71	72.44	73.58	74.55
	CNN RNN [24]	78.83	74.99	76.12	77.18
Skeletal	Hierarchical RNN [8]	84.36	79.03	81.52	82.92
	CNN LSTM [10]	87.11	82.37	85.05	86.03
	Temporal Sliding LSTMs [15]	87.94	81.55	85.22	86.07
	Visual Attention [18]	89.36	84.05	86.85	88.96
RGB + Depth	CNN LSTM	93.32	88.42	90.25	91.42

RGB and depth data across all subjects. The aim of this section is to investigate the suitability of RGB and depth information for action classification through deep learning networks. Given that, we compare the mRA from multiple architectures on three multi modal action data. Table III presents the results of our investigation. The networks were borrowed from previous methods and were trained from scratch on the datasets used in this work. All the networks are trained and tested only once. From Table III, we were able to generate two insights regarding the performance of the action recognition models. One is based on the use of input data and the other is on the deep networks. We see that RGB based methods performed poorly when compared to the other two modalities, depth and skeletal. This is because of the background noise that exists in the RGB video frame that are difficult to learn during training of spatial networks. Contrastingly, this background noise is relatively less in-depth frames, and it is completely absent in skeletal data. Hence, skeletal action recognition is the popular choice for producing higher accuracies with deep networks. Despite their success the skeletal action data becomes noisy when there is a joint overlap during the action sequence producing ambiguous results.

Simultaneously, skeletal action data is represented as time series data which is perfectly characterized and discriminated using RNNs and LSTMs together. These networks have produced the highest recognition accuracies across all datasets. However, modelling RGB and Depth as time series data by extracting features and inputting those features to recurrent networks has shown to improve performance. However, the most obvious choice of combination is the skeletal data with either depth or RGB. The fusion with skeletal data has improved the discriminating confidence of the networks. The most suitable network architecture is the hybrid CNN LSTM which can extract spatial and temporal dynamics of the action data. Contrasting to the regular phenomenon, we applied RGB and depth modalities to CNN LSTM architecture to generate a highly discriminating feature vector for action recognition. Table III shows that our proposed method is on par with the existing state of the art models and in fact better than some of the existing models. All the models are tested with cross subject data. Finally, the last subsection evaluates the networks for cross data validation.

E. Cross Data Validation

This section shows the experimental evaluation of CLANet across datasets. We found some of the common actions across datasets and evaluated the performance of CLANet with separate training and testing data from multiple datasets. Incidentally, we trained the CLANet with our BVRCAction3D dataset and tested with same actions from another dataset. We used seven common actions across datasets. The results of this experiment were presented as mean recognition accuracy across these seven actions used for training and testing in Table IV. Here, the network has to fine tuned multiple times and the recognition rates obtained are averaged across multiple runs of the algorithm. Table IV shows that the proposed network has capabilities in evaluating cross data action recognition. Interestingly, we found that training with less noisy data could result in good recognition accuracies when compared to a noisy data training. The average recognition was around 65% across datasets with the proposed CLANet with RGB and depth input data.

Despite better performance by the hybrid CNN LSTM architecture across RGB D action datasets for recognition tasks, there are many challenges such as view invariance, cross data and occlusions that need attention. We found that it is difficult to achieve high degree of robustness for some complex actions from the existing deep learning frameworks. Moreover, deep networks are data intensive models and require a wide variety to provide actionable intelligence across action recognition platforms. Finally, more hybrid models with multiple levels of abstraction are required for designing deployable action recognition models.

V. Conclusion

In this paper, we have proposed a novel approach for recognizing RGB D action data. Specifically, our method involves training of a hybrid CNN LSTM multi stream network on multi modal data, RGB and depth videos. The CNN network is designed to extract spatial features from both RGB and depth action frames. Subsequently, bidirectional LSTM network is used to model the sequential information in the extracted multi modal features at the output of the CNN. The hybrid CLANet is trained and tested using our generated BVRCAction3D dataset and other benchmark datasets for recognition. The results conclude that the proposed network is capable of achieving higher average recognition rates of around 93.32% on our dataset and an average of 90.24% across all benchmark datasets.

References

[1] Jian-Fang Hu, Wei-Shi Zheng, Jiahui Pan, Jianhuang Lai, and Jianguo Zhang. Deep bilinear learning for RGB-d action recognition. In *Computer Vision – ECCV 2018*, pages 346–362. Springer International Publishing, 2018.

[2] Sunitha Ravi, Maloji Suman, P.V.V. Kishore, Kiran Kumar E, Teja Kiran Kumar M, and Anil Kumar D. Multi modal spatio temporal co-trained CNNs with single modal testing on RGB–d based sign language gesture recognition. *Journal of Computer Languages*, 52:88–102, jun 2019.

[3] Yen-Yu Lin, Ju-Hsuan Hua, Nick C. Tang, Min-Hung Chen, and Hong-Yuan Mark Liao. Depth and skeleton associated action recognition without online accessible RGB-d cameras. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2014.

[4] Qiuxia Wu, Zhiyong Wang, Feiqi Deng, Zheru Chi, and David Dagan Feng. Realistic human action recognition with multimodal feature selection and fusion. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 43(4):875–885, jul 2013.

[5] Z. Gao, S.H. Li, Y.J. Zhu, C. Wang, and H. Zhang. Collaborative sparse representation leaning model for RGBD action recognition. *Journal of Visual Communication and Image Representation*, 48:442–452, oct 2017.

[6] Jing Zhang, Wanqing Li, Philip O. Ogunbona, Pichao Wang, and Chang Tang. RGB-d-based action recognition datasets: A survey. *Pattern Recognition*, 60:86–105, dec 2016.

[7] D. Srihari, P. V. V. Kishore, E. Kiran Kumar, D. Anil Kumar, M. Teja Kiran Kumar, M. V. D. Prasad, and Ch. Raghava Prasad. A four-stream ConvNet based on spatial and depth flow for human action classification using RGB-d data. *Multimedia Tools and Applications*, 79(17-18):11723–11746, jan 2020.

[8] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2015.

[9] E. Kiran Kumar, P. V. V. Kishore, A. S. C. S. Sastry, M. Teja Kiran Kumar, and D. Anil Kumar. Training CNNs for 3-d sign language recognition with color texture coded joint angular displacement maps. *IEEE Signal Processing Letters*, 25(5):645–649, may 2018.

[10] Juan C. Núñez, Raúl Cabido, Juan J. Pantrigo, Antonio S. Montemayor, and José F. Vélez. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. *Pattern Recognition*, 76:80–94, apr 2018.

[11] P. V. V. Kishore, D. Anil Kumar, A. S. Chandra Sekhara Sastry, and E. Kiran Kumar. Motionlets matching with adaptive kernels for 3-d indian sign language recognition. *IEEE Sensors Journal*, 18(8):3327–3337, apr 2018.

[12] Eepuri Kiran Kumar, P. V. V. Kishore, Maddala Teja Kiran Kumar, Dande Anil Kumar, and A. S. C. S. Sastry. Three-dimensional sign language recognition with angular velocity maps and connived feature ResNet. *IEEE Signal Processing Letters*, 25(12):1860–1864, dec 2018.

[13] Earnest Paul Ijjina and Krishna Mohan Chalavadi. Human action recognition in RGB-d videos using motion sequence information and deep learning. *Pattern Recognition*, 72:504–516, dec 2017.

[14] Anne Veenendaal, Elliot Daly, Eddie Jones, Zhao Gang, Sumalini Vartak, and Rahul S Patwardhan. Sensor tracked points and hmm based classifier for human action recognition. *Computer Science and Emerging Research Journal*, 5:4–8, 2016.

[15] Inwoong Lee, Doyoung Kim, Seoungyoon Kang, and Sanghoon Lee. Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017.

[16] Danilo Avola, Marco Cascio, Luigi Cinque, Gian Luca Foresti, Cristiano Massaroni, and Emanuele Rodola. 2-d skeleton-based action recognition via two-branch stacked LSTM-RNNs. *IEEE Transactions on Multimedia*, 22(10):2481–2496, oct 2020.

[17] Jiajia Luo, Wei Wang, and Hairong Qi. Spatio-temporal feature extraction and representation for RGB-d human action recognition. *Pattern Recognition Letters*, 50:139–148, dec 2014.

[18] Zhengyuan Yang, Yuncheng Li, Jianchao Yang, and Jiebo Luo. Action recognition with spatio–temporal visual attention on skeleton image sequences. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2405–2415, aug 2019.

[19] Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang. Real-time action recognition with enhanced motion vector CNNs. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016.

[20] Myunggi Lee, Seungeui Lee, Sungjoon Son, Gyutae Park, and Nojun Kwak. Motion feature network: Fixed motion filter for action recognition. In *Computer Vision – ECCV 2018*, pages 392–408. Springer International Publishing, 2018.

[21] Pichao Wang, Wanqing Li, Zhimin Gao, Chang Tang, and Philip O. Ogunbona. Depth pooling based large-scale 3-d action recognition with convolutional neural networks. *IEEE Transactions on Multimedia*, 20(5):1051–1061, may 2018.

[22] Weiyue Wang and Ulrich Neumann. Depth-aware CNN for RGB-d segmentation. In *Computer Vision – ECCV 2018*, pages 144–161. Springer International Publishing, 2018.

[23] Zhiyuan Shi and Tae-Kyun Kim. Learning and refining of privileged information-based RNNs for action recognition from depth sequences. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jul 2017.

[24] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with rCNN. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, dec 2015.

[25] Chao Li, Zhongtian Bao, Linhao Li, and Ziping Zhao. Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition. *Information Processing & Management*, 57(3):102185, may 2020.

[26] Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and Xiangyang Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *Proceedings of the 23rd ACM international conference on Multimedia - MM -15*. ACM Press, 2015.

[27] Lei Wang, Yangyang Xu, Jun Cheng, Haiying Xia, Jianqin Yin, and Jiaji Wu. Human action recognition by learning spatio-temporal features with deep neural networks. *IEEE Access*, 6:17913–17922, 2018.

[28] Chuankun Li, Pichao Wang, Shuang Wang, Yonghong Hou, and Wanqing Li. Skeleton-based action recognition using LSTM and CNN. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, jul 2017.

[29] Xuanhan Wang, Lianli Gao, Jingkuan Song, and Hengtao Shen. Beyond frame-level CNN: Saliency-aware 3-d CNN with LSTM for video action recognition. *IEEE Signal Processing Letters*, 24(4):510–514, apr 2017.

[30] Wen-Nung Lie, Anh Tu Le, and Guan-Han Lin. Human fall-down event detection based on 2d skeletons and deep learning approach. In *2018 International Workshop on Advanced Image Technology (IWAIT)*. IEEE, jan 2018.

[31] Zhiyuan Shi and Tae-Kyun Kim. Learning and refining of privileged information-based RNNs for action recognition from depth sequences. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jul 2017.

[32] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+d: A large scale dataset for 3d human activity analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016.

[33] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2012.

[34] Lu Xia, Chia-Chih Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, jun 2012.

TABLE IV. TABLE SHOWS MRA OF MULTI-SOURCE TRAINING AND TESTING OF THE PROPOSED CLANET

Training Dataset	Testing Dataset	Recognition Rates on Different Actions							
		Walking	Jumping	Jogging	Kicking	Hand Waving	Clapping	Running	Average Recognition
MSRDailyAction3D	BVRCAction3D	67.26	69.23	69.51	69.93	67.54	69.23	68.73	68.77
	MSRDailyAction3D	86.98	88.88	87.54	89.79	88.30	88.73	87.55	88.25
	UT Kinect	66.90	65.85	67.26	67.37	67.35	66.43	65.60	66.68
	NTU RGB D	69.52	61.00	69.36	60.67	60.25	61.52	69.00	65.18
BVRCAction3D	BVRCAction3D	98.85	97.55	97.26	96.81	97.51	96.97	97.50	97.49
	MSRDailyAction3D	68.94	68.50	66.35	66.90	66.53	67.89	67.33	67.49
	UT Kinect	65.89	64.73	65.86	65.61	65.13	65.25	65.73	65.45
	NTU RGB D	65.93	65.81	64.89	65.94	63.79	64.78	64.74	65.12
UT Kinect	BVRCAction3D	52.59	53.98	53.70	55.51	53.90	55.66	53.68	54.14
	MSRDailyAction3D	59.90	57.75	55.97	66.83	56.29	54.38	54.94	56.58
	UT Kinect	89.90	87.75	85.97	86.83	86.29	84.38	84.94	86.58
	NTU RGB D	58.97	58.75	59.90	57.95	50.46	50.75	50.96	55.67
NTU RGB D	BVRCAction3D	66.64	64.93	66.56	64.89	66.91	65.92	65.45	65.9
	MSRDailyAction3D	62.75	63.97	63.80	63.93	65.41	64.65	63.94	64.06
	UT Kinect	66.96	66.44	66.82	66.62	65.74	66.43	65.43	66.34
	NTU RGB D	94.80	94.90	94.94	94.80	93.53	93.65	93.58	94.31

[35] Li Liu, Ling Shao, and Peter Rockett. Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition. *Pattern Recognition*, 46(7):1810–1818, jul 2013.

[36] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

A Rule-based Approach toward Automating the Assessments of Academic Curriculum Mapping

Abdullah Alshanqiti[1]
Tanweer Alam[2], Mohamed Benaida[3], Abdallah Namoun[4], Ahmad Taleb[5]
Faculty of Computer and Information Systems
Islamic University of Madinah
Madinah, 42351, Saudi Arabia

***Abstract*—Curriculum mapping is the blueprint of a successful academic program. It is progressively utilised in higher education as a monitoring tool in the current age of standard-based regulations and empowers program leaders and course instructors to align their curricula for the offered courses and the corresponding learning outcomes. It is often depicted by a two-dimensional matrix expressing the relationship between the students learning outcomes (i.e., SOs) and the courses. However, its mapping remains a challenging exercise, even for experienced program leaders. The complexity stems from the fact that mistakes are prone to happen during the mapping, and program leaders need to be aware of the rules and the acceptable practices of curriculum-effective mapping. Besides, it is not straightforward to spot contradictions in the SO-course mappings. Consequently, this paper aims to tackle these challenges by investigating effective-mapping rules from existing curriculum mappings, which allows one to inspect the SO-course mappings, discover inefficiencies, and provide suggestions for improving the curriculum mapping. We identify the main mapping criteria and propose a rule-based algorithm for curriculum matrix assessments. This algorithm is implemented in an online application and evaluated using a user-based experiment, relying on curriculum mapping experts. The findings have shedded light on the promise of our approach.**

***Keywords*—Rule based algorithm; curriculum mapping; student learning outcomes; program outcomes; assessment; curriculum matrix**

I. Introduction

Curriculum mapping is the blueprint of a successful academic program. It is progressively utilised in higher education as a monitoring tool in the current age of standard-based regulations and empowers program leaders and course instructors to align their curricula for the offered courses and the corresponding learning outcomes [1]. Curriculum mapping approaches are implemented in academic institutions in a bid to improve the quality of education [2], culture of involvement, and cooperation. Such approaches can assist students in attaining the learning outcomes of educational programs more effectively and highlight the deficiencies within these programs [3], [4]. Curriculum mapping refers to the process of regulating program learning outcomes with the offered courses to recognize or highlight educational discrepancies, inefficiencies, as well as misalignments with a perspective program outcome [5]. The results include improvements to the overall consistency of course-program mappings and acceptable coverage of the program student outcomes (i.e. SOs) [6]. Moreover, curriculum mapping explores how well and to what extent a course instructor might have represented the contents to cover the academic requirements described by the student learning outcomes [7]. Curriculum mapping is often documented for every semester, throughout all subjects, and by each lecturer directly once the class's material is designed.

Despite its evident benefits, curriculum mapping remains a challenging exercise, even for experienced program leaders. The complexity stems from the fact that mistakes can be easily made during the mapping. Moreover, program leaders need to be aware of the rules and the acceptable practices of curriculum-effective mapping. Furthermore, manual mapping involves course instructors, which makes it resource intensive. Besides, it is not straightforward to spot contradictions in the SO-course mappings [8]. Consequently, this paper aims to tackle these challenges by investigating effective-mapping rules from existing curriculum mappings, which allows one to inspect the SO-course mappings, discover inefficiencies, and provide suggestions for improving the curriculum mapping.

This paper focuses on developing an effective relationship between the courses and program learning outcomes, represented as a two-dimensional matrix, in an automated model. Quality assurance and accreditation agencies mandate that curriculum courses must be consistent with the program's educational goals and clearly state that curriculum mapping should concentrate on the fulfillment of each program/student outcome. More in detail, this paper is motivated by the following two research questions: (**Question One**) what are the best practices to achieve effective curriculum mapping in educational programs? and (**Question Two**) how can we implement these practices as part of a rational mapping solution?. In this paper, aiming at assessing the curriculum matrix of academic programs besides giving a set of recommendations for improving course- alignments, we make two contributions based on our observations on several existing curriculum mappings.

- We identify the main rules (criteria) for assessing curriculum matrix based on studying several well-defined curriculum mappings.
- We propose a rule-based algorithm for curriculum matrix assessments. This algorithm is implemented and encapsulated in a web-based application that is publicly available online.
- We report on a used-based evaluation (i.e., based on curriculum mapping experts) for validating our proposal.

The remainder of the paper is divided into four sections. Section II represents the related works, shedding light on the well-used mapping techniques for curriculum evaluation. Section III introduces our rule-based algorithm for assessing the curriculum mapping, besides discussing the derivation of these rules from the existing curriculum matrix. Section IV describes the proposed online assessment tool. Section V presents and discusses the conducted user-study for evaluating our proposal, and finally, Section VI concludes the paper with directions for future research.

II. Related Work

In this section, we review several types of relevant work, categorized into three aspects. First, we discuss a few strategic design approaches for course-alignments with program outcomes in two-dimensional matrix representation. Analyzing these approaches allowed us to derive the core mapping rules (or criteria). In the second category, we focus particularly on analytic mapping techniques for curriculum evaluation. Finally, since this paper attempts to assess the quality of a given curriculum matrix along with its set of program outcomes, we review the closely related approaches that focus on alignment's quality between courses and program outcomes.

A. Strategic Design for Creating a Curriculum Matrix

In [9], the authors investigate how curriculums have developed in Estonia since the late 90s where the government took the initiative to reform national school curriculums. A comparison is made between their curriculum and that of Great Britain and Latvia while focusing on the historical and theoretical aspects. Plaza et al. [10] proposed a curriculum mapping approach, including program evaluation. In their article, the authors represent the association between the graphical curriculum maps of students and teachers concerning the ranking of the relative emphasis of each domain, indicating the adjustment between the curricula intended/delivered and obtained [10].

Spencer, et al. [11] have presented curriculum mapping as an embedded tool with graduate capabilities. In this article, the authors discussed a method for collecting, analyzing, and presenting current teaching and graduate capacity evaluation information. Their conceptual approach promotes interactive curriculum development exercise, whereas the resulting graphs continue providing conceptual representations of present procedures and measurements of where curriculum redesign should be positioned. Lacerda and Sepel [12] research curriculum perceptions through educators collect information regarding their experience and come up with ideas that will positively change the curriculum. Their results identified that post-critical theories were more accepted than critical theories, whereas the results regarding traditional theories were not clear enough. Nevertheless, it was heavily linked with the organization of classroom practices.

Linden, et al. [13] describe the influence of curriculum theories in higher education while concentrating on their historical and conceptual roots. Results show that adapting modern higher education environments requires identifying the differences between normative and critical curriculum while keeping in mind the curricula' influence. Roth and Thom [14] clarify the differences in curriculum theory over the past decade or two; they claim that number of curriculum topics studied has increased along with methodological and theoretical expansions. Their research indicates that curriculum inquiry is becoming more widespread and that there is potential for a cross-cultural attitude. Furthermore, it explores the relationship between the perception of the structure of the field and curriculum inquiry.

Sweden implemented the new curriculum in 2011, which was investigated by [15]. This investigation indicates that the teachers' curriculum agency identifies three spaces in an educational classroom: interactive space, collective space, and individual space. The teachers' feedback displayed clear evidence that the change in the curriculum required a greater level of content assessment, whereas the teachers can more easily assess abilities as they may be required to get involved with the students directly. Ghaderia [16] study ideas regarding peace-based curriculum, in a time where they believe peace is crucial for global stabilization with the technological warfare advancements. They believe that through merging post-modern theories and liberal theories (i.e., by merging the similarities and differences between the two), the ultimate peace-based curriculum can be achieved.

B. Analytic Mapping Techniques for Curriculum Evaluation

Treadwell, et al. [17] explores the launch of curriculum mapping on the web-based interactive learning opportunities, objectives, and outcome platform (LOOOP) via interviewing 30 lecturers concerning their experience with this curriculum mapping. Overall, the results showed that although the participants did not immediately master the use of this system and did face some slow development issues, except that they believed it was ultimately beneficial to the learning program and the students within the course due to several reasons, including; communication, usability, and transparency.

According to George Mason University's Health Administration Program experience, Perlin [18] examined the curriculum mapping for a program evaluation. Their examination helped in establishing a framework with the merging of an enhanced analytic process and a mapping exercise. This framework was initiated by identifying the analytic and technical methods by investigating the mapping process used within the university and its overall course objectives. As a result, setbacks were identified and pointed out so that it can be adjusted in the future, and users are aware of how to overcome these hurdles. Therefore users of this curriculum-mapping program will play a role in upgrading the quality of the university graduate program.

Avella, et al. [19] claim that although the use of learning analytics in higher education is very promising, except that one of the flaws associated with the immediate introduction of learning analytics in this sector is that the users are not able to make use of this system effectively or to its maximum potential. Therefore, the researchers studied the techniques that learning analytics implement such as visual data analytic techniques, relationship mining, and so on. Consequently, the benefits and the challenges of learning analytics were identified and listed according to its use in higher education, therefore educators can make use of this implemented system successfully and effectively which in turn would improve the quality of teaching within higher education.

Pat Hutchings [20] has published a report on the alignment of educational outcomes and practices. This research explores and disseminates ways in which educational programs and organizations could use evaluation information effectively to inform and enhance academic education and interact indirectly with decision-makers. Yates and Millar [21] discover the physics curriculum, in particular within Australian universities and schools. Physics was specifically chosen as it is a science that must adapt to change and update as time goes by. Their focus was to investigate whether or not the curriculum can be logically derived depending on the discipline.

However, Ghaith Al-Eyd1 et al. [22] claim that the use of curriculum mapping in a new medical school help provide a better understanding of the system as compared to other medical schools as well as the educational environment. Their research identified which factors benefitted from the use of their curriculum mapping; it mainly involved improving the level of organization and providing more evident learning outcomes that demonstrate the links between the entire course and outcomes.

C. *Quality of Course-Outcomes Alignments*

In 2019, Buker and Niklason [23] presented an improved curriculum mapping model. They tried to include some essential standards to help develop an assessment process. The recommended approaches involve assessing the program's mission and evaluating the course outcomes based on the criteria. Uchiyama and Radin [24] published an article on curriculum mapping in higher education. They have represented curriculum mapping as a mechanism that generates a visualization of the curriculum based on real-time data to improve the quality of education.

Lam and Tsui [25], published an article on the examination of alignments in curriculum mapping. This research proposes that curriculum mapping can be a helpful tool for assessing how the students approved learning outcomes are aligned with the classes provided by the academic faculty. M Jacobsen, et al [26] presented the article on action research for graduate program improvements. They have presented continuous program improvement techniques derived from actions and suggestions, which emerged from a year-long faculty-led, institutionally approved evaluation curriculum of academic programs. Jacobsen, et al. [27] has identified several factors that contribute to improving the program at hand. These factors include interconnecting the courses efficiently and effectively, providing a greater level of ethical support to students within the graduate program, etc. This research has helped identify the strengths of this graduate program and methods to fix its flaws.

The original contributions of this paper that differ from all previous works lies in examining various curriculum-mappings for rules extraction in addition to our proposed rule-based algorithm for mapping assessments. We will discuss these contributions in the next section.

III. Assessing Curriculum Matrix

In this section, we present our rule-based procedure for assessing the curriculum mapping based on students learning outcomes. Given an academic program consisting of a set of outcomes, the proposed procedure assesses its two dimensional matrix that links the top-level outcomes with the low-level course-outcomes. More significantly, our procedure can help to give insights into curriculum mappings design, pinpointing the main criteria used for quality measurements. In order to automate our procedure, we have studied many different curriculum mappings, attempting to extract common rules (i.e., rules potentially applied during the development of the mappings) as a bottom-up approach. To this end, we consider a three-phase methodology to achieve the objectives of this paper:

- Identify the existing curriculum mapping criteria and requirements.
- Propose (and give an implicit implementation) of an algorithm to measure the quality of the curriculum mapping.
- Evaluate the effectiveness of the proposed procedure based on surveys conducted with curriculum experts (discussed in the next section).

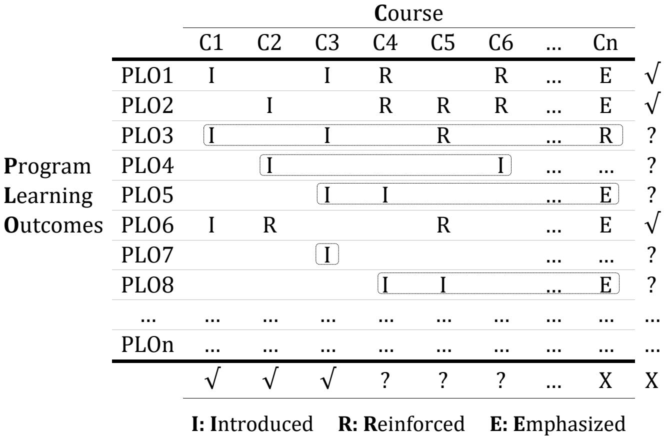
For the first phase, the Faculty of Computer and Information Systems (FCIS) at the Islamic University of Madinah was selected as a research site with a four-year history of curriculum mappings. In addition, we assume standard mapping letters (I, R and E) when designing curriculum mappings (i.e., link courses to program learning outcomes). These mapping letters are broadly employed in various academic accreditation agencies (e.g., ABET, AACSB, NCAAA) with the following common purposes:

- Introduced (I): It means that students are not expected to be proficient with the content or expertise. Learning activities here focus on basic knowledge, comprehension, skills, and competencies at an entry-level (typically assigned in first to second-year courses).
- Reinforced (R): Expecting students to have a necessary amount of knowledge and understanding of the content or talents. Learning activities concentrate on enhancing and strengthening knowledge, skills, and expanding complexity (typically assigned in Second, Third, Fourth-year courses).
- Emphasized (E): Here, students are expected to have a robust as well as a sophisticated understanding, expertise, or competency base. Instructional and learning activities focus on using the content or skills in multiple contexts and at multiple levels of complexity (capstone courses).

Figure 1 exemplifies a representation of the curriculum matrix, which illustrates different issues, such that not all PLOs are properly aligned with courses in accordance with the standard rules for creating a robust mapping. In a nutshell, to address such issues, the following designing steps are suggested to obtain a proper curriculum mapping that may assist in getting more accurate learning-outcome measurements.

1) *Learning Outcomes*: The learning outcomes at the program level, as well as the course level, must be defined. The learning outcomes must be consistent with the objectives, goals, and mission of the program.
2) *Mapping*: the learning outcomes of the courses must be aligned with the program learning outcomes. The prerequisite courses should be taken into consideration to ensure cumulative and consecutive learning achievement of students.

3) *Alignment Criteria*: a set of criteria should be applied to build curriculum mapping. For instance, the logical order of the alignment levels (I, R, E), the number of courses aligned to each program, etc.
4) *Involvement of stakeholders*: the faculty members, other educational expertise, and the curriculum developer should be involved in developing the curriculum mapping.



		Course							
		C1	C2	C3	C4	C5	C6	...	Cn	
	PLO1	I		I	R		R	...	E	√
	PLO2		I		R	R	R	...	E	√
	PLO3	I		I		R		...	R	?
Program	PLO4		I				I	...	...	?
Learning	PLO5			I	I			...	E	?
Outcomes	PLO6	I	R			R		...	E	√
	PLO7			I				...	...	?
	PLO8				I	I		...	E	?
	...	...	...	...	...	...	...	...	...	...
	PLOn	...	...	...	...	...	...	...	...	...
		√	√	√	?	?	?	...	X	X

I: Introduced
R: Reinforced
E: Emphasized

Fig. 1. An Exemplified Representation of a Poor-Defined Curriculum Matrix

TABLE I. A MODEL OF CURRICULUM MATRIX

	Introductory	Required Courses					Capstone
	Courses	RC1	RC2	RC3	–	RCn	Course
PLO-1/SO1	I	R	R		–	R	E
PLO-2/SO2	I			R	–	R	E
PLO-3/SO3	I		R	R	–		E
PLO-4/SO4	I			R	–	R	E
PLO-5/SO5	I	R			–	R	E
PLO-6/SO6	I			R	–	R	E
PLO-n/SOn	I	R	R	R	–	–	E

In Table I, we illustrate a curriculum matrix as a model that visually represents the alignment between program learning outcomes and the curriculum courses. Broadly speaking, there exist four recommend rules of thump to apply for measuring such a curriculum matrix's quality. These rules are derived from the best quality assurance practices as well as from the well-known accreditation agencies [28], [29], [30]. More in detail, the first rule suggests that all course outcomes (with I, R, E) should be aligned to at least three courses [28]. While the second rule focuses on not to align many courses to a particular outcome. Here, the recommended number of courses covering a particular outcome is 3-5 [29]. The third rule suggests attaining the learning outcomes in a logical order, which empowers students to gradually progress their attainments from the first level to greater experience levels. (I, R, E) [30]. The fourth rule suggests having (I) in only the first and second academic levels, which typically should cover all introductory courses. Similarly, (R) is recommended to cover all the mid-advanced courses (i.e., from the third academic level to the last level), and (E) for only capstone courses such as the final senior project [30]. In this paper, we consider these four abstracted rules as the core criteria for assessing a given curriculum matrix using our implemented algorithm, introduced explicitly in algorithm 1.

Algorithm 1: Rule-based procedure for curriculum matrix assessments

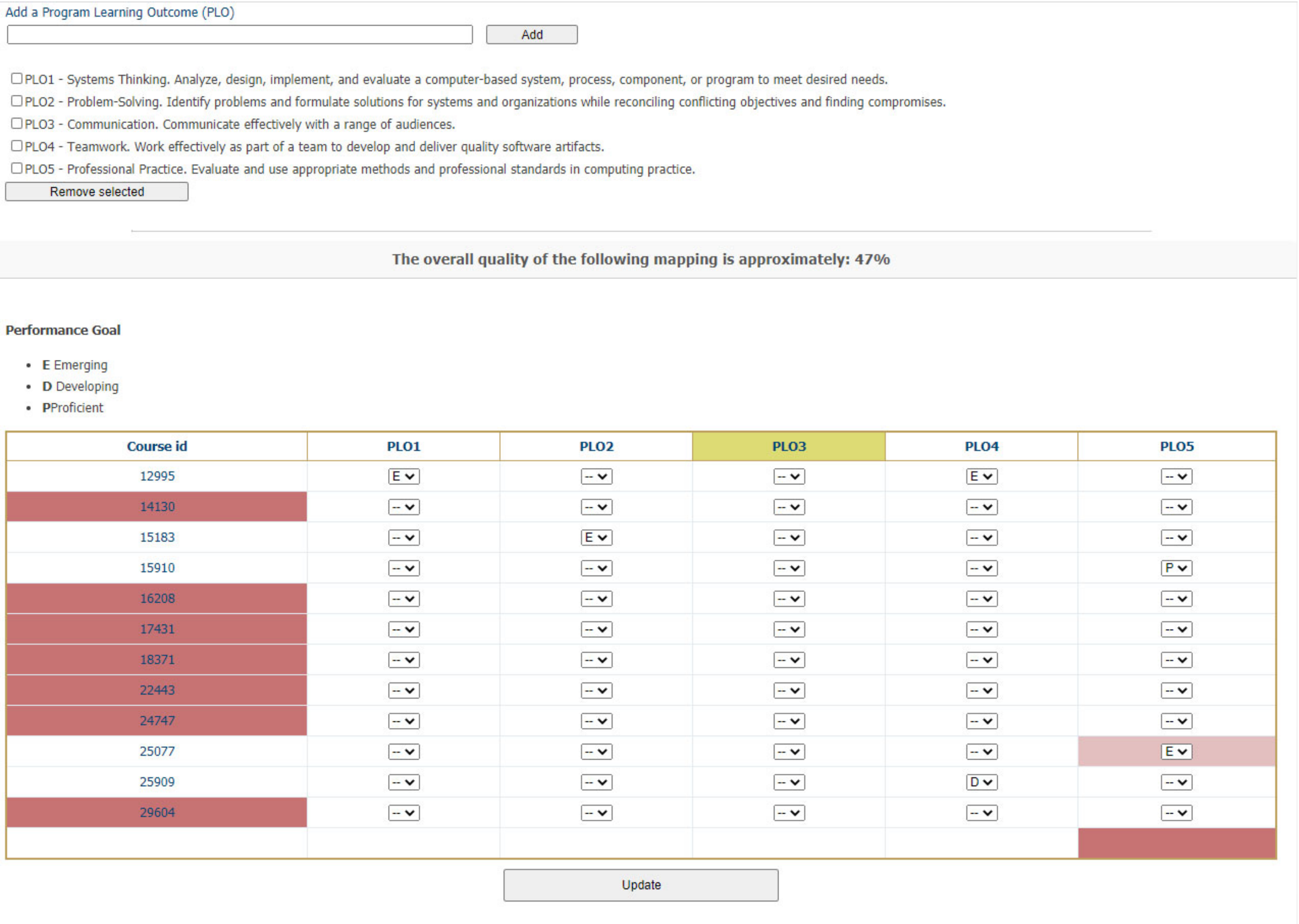
```
input : $CM = [c, o]$: a curriculum matrix with $c$
        courses and $o$ program outcomes.
output: $Rec$: a set of recommendations, and $Acc$: the
        estimated accuracy of the overall mapping.

$Rec \leftarrow \emptyset$
$iCount \leftarrow 0$

for $i \leftarrow 1$ to $|c|$ do
    for $j \leftarrow 1$ to $|o|$ do
        $m \leftarrow CM[i, j]$  Getting the mapping letter
          (i.e., I, R, E, or $\emptyset$)
        $re \leftarrow$ RuleValidation($CM$, $m$)
          Validating the mapping letter using our
          identified four rules
        if $\neg re == \emptyset$ then $iCount += 1$
        $Rec$.append($re$)

$Acc = \frac{iCount}{|c| \cdot |o|}$
return $Rec$, $Acc$

Function RuleValidation($CM$, $m$)
    $re[] \leftarrow \emptyset$
    for $k \leftarrow 1$ to 4
        Here we perform another two nested loops over
        $CM$ for rule validations
        if isRuleMatched('Rule[k]',m, $CM$) then
            $re$.append(getRecommend('Rule[k]', $m$))
    return $re$
```

IV. Web-based Application for Assessing Curriculum Mapping

This section presents our proposed helper web-based application tool for automating the assessments of a given academic program. It is publicly available at (http://iudss.com/). The curriculum mapping feature is implemented based on algorithm 1, which provides different runtime figures/charts to explore the efficiency of the mapping at both program and student levels. We imagine that giving insights about the generated figures can help in understanding the overall learning outcomes and efficiency of the educational program at hand. In a nutshell, our tool produces several dynamic charts with the percentage of the mapping quality. An example of these charts is illustrated in Figure 2. Here, a distinct color scheme is used to signify the mapping issues, relying on the identified validating rules.

V. Preliminary Empirical Results and Discussion

A straightforward technique to gauge the efficacy of designing a curriculum mapping is to consider various sources, such as students, universities, disciplines, and higher education institutions. Almost every stakeholder has their expectations as to how the curriculum impacts current and future outcomes. In this research, we investigate through a simple user-study (i.e., conducted with curriculum mapping experts) the validity



Add a Program Learning Outcome (PLO)

☐ PLO1 - Systems Thinking. Analyze, design, implement, and evaluate a computer-based system, process, component, or program to meet desired needs.
☐ PLO2 - Problem-Solving. Identify problems and formulate solutions for systems and organizations while reconciling conflicting objectives and finding compromises.
☐ PLO3 - Communication. Communicate effectively with a range of audiences.
☐ PLO4 - Teamwork. Work effectively as part of a team to develop and deliver quality software artifacts.
☐ PLO5 - Professional Practice. Evaluate and use appropriate methods and professional standards in computing practice.

The overall quality of the following mapping is approximately: 47%

Performance Goal

- E Emerging
- D Developing
- PProficient

Course id	PLO1	PLO2	PLO3	PLO4	PLO5
12995	E	--	--	E	--
14130	--	--	--	--	--
15183	--	E	--	--	--
15910	--	--	--	--	P
16208	--	--	--	--	--
17431	--	--	--	--	--
18371	--	--	--	--	--
22443	--	--	--	--	--
24747	--	--	--	--	--
25077	--	--	--	--	E
25909	--	--	--	D	--
29604	--	--	--	--	--

Fig. 2. Screenshot of the Proposed Tool Illustrating the Quality of the Curriculum Mapping Matrix.

of our proposed method compared with various practices for designing curriculum mapping. At the end of this section, we will identify the main benefits that one can gain by having well-defined curriculum mapping.

We have extracted the opinions of curriculum mapping experts, academic professionals, and senior faculties through questionnaires. Commonly, there are many data-collection methods to consider; however, this paper's authors decided to implement a comprehensive questionnaire as it is the most natural and reliable method. This method provides valuable information regarding the experience of the participants on a specific system and measures their level of satisfaction [31], [32], [33]. We have gathered seven experts for completing our questionnaire that consists of ten questions. The first question aims to identify the expertise field of the involved experts, see Table II. The remaining nine questions concentrate on curriculum mapping, scaled as follows; very satisfied, satisfied, neutral, dissatisfied, and finally very dissatisfied.

Most of the experts involved in this questionnaire are curriculum experts (42.9%), whereas the remaining experts are academic professionals (28.6%) and senior faculty members (28.6%). Their valuable experience and knowledge allowed them to analyze how helpful the intelligent mapping curriculum evaluated was. More than half claimed that they were very satisfied with what they were provided with (42.9% and 14.3%, respectively). None of the experts were very dissatisfied; however, 14.3% were dissatisfied with how helpful the intelligent mapping curriculum was, whereas 28.6% of the experts took a neutral standpoint. In the third part of the questionnaire, we asked the experts ... *Do you think curriculum mapping will be helpful for the satisfaction of learning outcomes?* and only 14.3% of them were dissatisfied and another 14.3% were neutral. The remaining experts were either very satisfied or satisfied (42.9% and 28.6% respectively).

The expert's opinions regarding whether or not intelligent curriculum mapping will evaluate the weaknesses of curriculum mappings is as follows; none of the experts showed a negative view in regards to this question, but only 14.3% of them were neutral and the remaining showed a positive attitude towards this question. The fifth question states ... *Do you think that the curriculum mapping could visibly allow the curriculum and illustrate the connections between the courses and the student's learning outcomes?*. The results of this question varied, 57.1% were very satisfied, 28.6% were neutral and the remaining 14.3% were dissatisfied. The next question discusses whether intelligent curriculum planning will generate the recommendations for improvements in curriculum mapping. Once again, there was no negative responses to this question, but 14.3% were neutral and the remaining had a positive response (42.9% for both very satisfied and satisfied). Question seven debates the transparency of intelligent curriculum mapping; 42.9% were very satisfied, 28.6% were satisfied and the remaining were neutral.

TABLE II. RESULTS OF THE QUESTIONNAIRE CONTAINING TEN QUESTIONS (Q1-Q10)

	Expert								
Question	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	Total (# 35)	Average
Q2	4	5	5	5	3	2	3	27	3.9
Q3	4	5	5	5	3	2	4	28	4.0
Q4	4	5	5	5	3	4	4	30	4.3
Q5	5	5	5	5	3	2	3	28	4.0
Q6	4	5	5	5	3	4	4	30	4.3
Q7	4	5	5	5	3	4	3	29	4.1
Q8	4	5	5	5	3	4	4	30	4.3
Q9	4	4	5	5	3	4	3	28	4.0
Q10	4	5	5	5	3	4	4	30	4.3

Q1: State what describe you the best from the following?

Expert **1**	Curriculum Expert
Expert **2**	Senior Faculty Member
Expert **3**	Senior Faculty Member
Expert **4**	Academic Professional
Expert **5**	Curriculum Expert
Expert **6**	Curriculum Expert
Expert **7**	Academic Professional

Q2: How helpful is the intelligent curriculum mapping provided to you?
Q3: Do you think the curriculum mapping will be helpful for the satisfaction of learning outcomes?
Q4: Do you think intelligent curriculum mapping will evaluate the weaknesses of curriculum mappings?
Q5: Do you think that the curriculum mapping could visibly allow the curriculum and illustrate the connections between the courses and the student's learning outcomes?
Q6: Do you think that the Intelligent curriculum mapping will generate recommendations for improvements in curriculum mapping?
Q7: Do you think that the Intelligent curriculum mapping will be transparent?
Q8: Do you think that the Intelligent curriculum mapping will align all student learning outcomes fairly?
Q9: Do you think the Intelligent curriculum mapping provided the right amount of theoretical and practical experience?
Q10: Please set your level of satisfaction for the intelligent curriculum mapping?

The eighth question states ... *Do you think that the intelligent curriculum planning will align all student learning outcomes fairly*. Here, most experts showed a positive reaction to this question, where 42.9% were very satisfied and another 42.9% were satisfied, the remaining 14.3% took a neutral standpoint. The next question inquires the experts about their opinions regarding the level of theoretical and practical experience of the intelligent curriculum mapping; 28.6% were very satisfied, 42.9% claimed they were satisfied and the remaining 28.6% had a neutral viewpoint. Finally, the results of the level of satisfaction of the experts were gathered, which showed that there were no negative responses, 14.3% were neutral and the remaining experts all had a positive outlook (85.8% of the experts were satisfied or very satisfied) on the level of satisfaction for the intelligent curriculum mapping. In Table II, we present the results of each question for each expert besides the calculated average for each question.

To summing up our observations, we list the main benefits of having an accurate representation of curriculum mapping, as follows:

- Ensuring the consistency of courses along with course outcomes.
- Effectively applied improvement techniques such as continuous, constant, standardized, and iterative.
- Help for continuous academic learning (consistency or incorporation).
- Revise and analyze learning outcomes).
- Enabling the professionalism of the academic staff (able to share the learning process).
- Discussing transparency problems (specifications, evaluation of programs, student support, and optimization of program outcomes),
- Criteria for quality assurance. With consideration for the program's curriculum, each course will need to create a plan to evaluate student learning outcomes.

TABLE III. STATISTICS OF LEARNING OUTCOMES AGAINST THE LEVEL OF COURSES

	Introduced	Reinforced	Emphasized
Disciplinary Knowledge	4	4	4
Critical Thinking	8	2	0
Communication	1	3	10
Research Skills	2	2	2
Ethical Reasoning	2	0	0

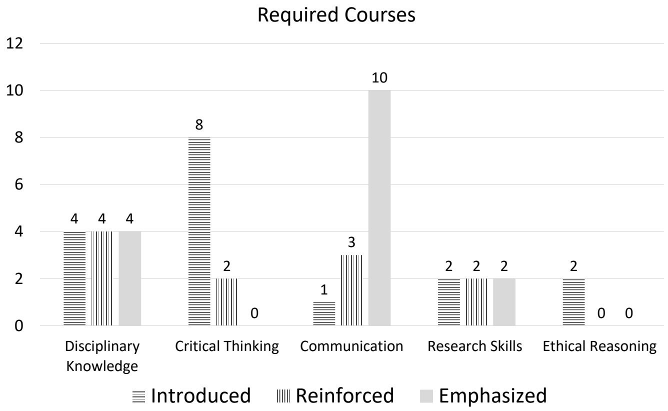


Fig. 3. Student's Learning Outcomes for *Required Courses*

The exhaustive assessment plan involves the learning outcomes, the process used to evaluate each outcome, the benchmark of each strategy, the person responsible for collecting the information, the information gathering regularity, the responsibility for interpreting the results and detecting modifications, as well as measuring enhancements, illustrated in Table III and Figure 3. We assume a total of 14 students' learning outcomes are divided into Disciplinary Knowledge, Critical Thinking, Communication, Research Skills, and Ethical Reasoning for the required courses [34]. According to the results of this study, most of the experts are satisfied with the proposed approach, see the results of the opinions of the experts in Table II. The curriculum review would be routine, but the number of variables, like input data and the facilities available, depends upon whether the entire curriculum is evaluated. Even so, the institutions must be aware of the need for a large-scale assessment and strategy to conduct the complete assessment. Whenever course outcomes are defined, it enables their mapping into the courses.

VI. Conclusion and Future Scopes

This paper has examined the curriculum matrix assessment using four criteria extracted from existing well-defined curriculum mappings. We have embedded these criteria into our proposed rule-based algorithm for assessments and recommendations. This algorithm observes the curriculum-mapping effectively and recommends potential improvements according to the identified rules (i.e. used to measure curriculum mapping quality). Conventionally, the curriculum mapping could aid in the creation and coordination of the consistency (developing, systematic, and goal-orientation) of the learning environment. At present, curriculum mapping approaches are often used by educational institutions to review and enhance their curriculum consistency toward attaining high level outcomes. In the future, we plan to extract more robust rules from existing well-defined curriculum mappings using Rule-based Machine Learning (RBML). In particular, we will be investigating the efficiency of using the regression by rule induction technique for spotting more complex issues in the curriculum matrix.

Acknowledgment

The authors acknowledge the Islamic University of Madinah, represented by the Deanship of Scientific Research, for the financial support of the Distinguished Project under the Grant $\#22 - 2018 - 2019$.

References

[1] P.-W. Chen and Y.-H. Wu, "Constructing a curriculum map to support learning navigation," in *2009 Joint Conferences on Pervasive Computing (JCPC)*. IEEE, 2009, pp. 45–50.

[2] M. Aljohani and T. Alam, "Design an m-learning framework for smart learning in ad hoc network of android devices," in *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. IEEE, 2015, pp. 1–5.

[3] A. Rahimi, S. A. M. Borujeni, A. R. N. Esfahani, and M. J. Liaghatdar, "Curriculum mapping: a strategy for effective participation of faculty members in curriculum development," *Procedia-Social and Behavioral Sciences*, vol. 9, pp. 2069–2073, 2010.

[4] A. M. Alshanqiti and A. Namoun, "Predicting student performance and its influential factors using hybrid regression and multi-label classification," *IEEE Access*, vol. 8, pp. 203 827–203 844, 2020.

[5] S. Farrell, C. Bodnar, and T. Forin, "Using concept mapping to develop inclusive curriculum," in *2017 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2017, pp. 1–3.

[6] H. S. Joyner, "Curriculum mapping: A method to assess and refine undergraduate degree programs," *Journal of Food Science Education*, vol. 15, no. 3, pp. 83–100, 2016.

[7] T. Herrmann and T. Leggett, "Curriculum mapping: Aligning content and design," *Radiologic technology*, vol. 90, no. 5, pp. 530–533, 2019.

[8] T. Alam and M. Aljohani, "M-learning: Positioning the academics to the smart devices in the connected future," *JOIV: International Journal on Informatics Visualization*, vol. 4, no. 2, 2020.

[9] V. Rouk, "From times of transition to adaptation: Background and theoretical approach to the curriculum reform in estonia 1987-1996." *Bulgarian Comparative Education Society*, 2013.

[10] C. M. Plaza, J. R. Draugalis, M. K. Slack, G. H. Skrepnek, and K. A. Sauer, "Curriculum mapping in program assessment and evaluation," *American Journal of Pharmaceutical Education*, vol. 71, no. 2, 2007.

[11] D. Spencer, M. Riddle, and B. Knewstubb, "Curriculum mapping to embed graduate capabilities," *Higher Education Research & Development*, vol. 31, no. 2, pp. 217–231, 2012.

[12] C. C. Lacerda and L. M. N. Sepel, "Basic school teachers' perceptions about curriculum theories," *Educação e Pesquisa*, vol. 45, 2019.

[13] J. Lindén, J. Annala, and K. Coate, "The role of curriculum theory in contemporary higher education research and practice," in *Theory and method in higher education research*. Emerald Publishing Limited, 2017.

[14] J. Linden, J. Annala, and K. Coate, "The role of curriculum theory in contemporary higher education research and practice," in *Theory and method in higher education research*. Emerald Publishing Limited, 2017.

[15] D. Alvunger, "Teachers curriculum agency in teaching a standards-based curriculum," *The Curriculum Journal*, vol. 29, no. 4, pp. 479–498, 2018.

[16] M. Ghaderia, "Peace-based curriculum based on the theories of difference and similarity," *Procedia-Social and Behavioral Sciences*, vol. 15, pp. 3430–3440, 2011.

[17] I. Treadwell, O. Ahlers, and G. Botha, "Initiating curriculum mapping on the web-based, interactive learning opportunities, objectives and outcome platform (looop)," *African Journal of Health Professions Education*, vol. 11, no. 1, pp. 27–31, 2019.

[18] M. S. Perlin, "Curriculum mapping for program evaluation and cahme accreditation," *Journal of Health Administration Education*, vol. 28, no. 1, pp. 33–53, 2011.

[19] J. T. Avella, M. Kebritchi, S. G. Nunn, and T. Kanai, "Learning analytics methods, benefits, and challenges in higher education: A systematic literature review." *Online Learning*, vol. 20, no. 2, pp. 13–29, 2016.

[20] P. Hutchings, "Aligning educational outcomes and practices. occasional paper# 26." *National Institute for Learning Outcomes Assessment*, 2016.

[21] L. Yates and V. Millar, "Powerful knowledge curriculum theories and the case of physics," *The Curriculum Journal*, vol. 27, no. 3, pp. 298–312, 2016.

[22] G. Al-Eyd, F. Achike, M. Agarwal, H. Atamna, D. N. Atapattu, L. Castro, J. Estrada, R. Ettarh, S. Hassan, S. E. Lakhan *et al.*, "Curriculum mapping as a tool to facilitate curriculum development: a new school of medicine experience," *BMC medical education*, vol. 18, no. 1, pp. 1–8, 2018.

[23] M. Buker and G. Niklason, "Curriculum evaluation & improvement model," *The Journal of Health Administration Education*, vol. 36, no. 1, p. 37, 2019.

[24] K. P. Uchiyama and J. L. Radin, "Curriculum mapping in higher education: A vehicle for collaboration," *Innovative Higher Education*, vol. 33, no. 4, pp. 271–280, 2009.

[25] B.-H. Lam and K.-T. Tsui, "Examining the alignment of subject learning outcomes and course curricula through curriculum mapping," *Australian Journal of Teacher Education*, vol. 38, no. 12, p. 6, 2013.

[26] M. Jacobsen, S. E. Eaton, B. Brown, M. Simmons, and M. McDermott, "Action research for graduate program improvements: A response to curriculum mapping and review," *Canadian Journal of Higher Education/Revue canadienne d'enseignement supérieur*, vol. 48, no. 1, pp. 82–98, 2018.

[27] M. Jacobsen, M. McDermott, B. Brown, S. E. Eaton, and M. Simmons, "Graduate students research-based learning experiences in an online master of education program," *Journal of University Teaching and Learning Practice*, vol. 15, no. 4, 2018.

[28] "Instructions: Curriculum map requirements." 2020. [Online]. Available: https://cte.tamu.edu/getattachment/Faculty-Teaching-Resource/Program-ReDesign/Curriculum-Mapping/Curriculum-Process-Overview-and-Instructions-SS-DF-(1).pdf.aspx?lang=en-US

[29] "Assessment and curriculum support center, university of hawai'i at manoa." 2020. [Online]. Available: http://manoa.hawaii.edu/assessment/howto/mapping.htm

[30] "Curriculum design, university of south florida." 2020. [Online]. Available: https://www.usf.edu/atle/teaching/curriculum-design.aspx

[31] E. Codó, L. Dans, and M. M. Wei, "Interviews and questionnaires," *The Blackwell guide to research methods in bilingualism and multilingualism*, pp. 158–176, 2008.

[32] M. Benaida *et al.*, "Developing arabic usability guidelines for e-learning websites in higher education," Ph.D. dissertation, University of Salford, 2014.

[33] M. L. Patten, *Questionnaire research: A practical guide*. Routledge, 2016.

[34] "Curriculum mapping: Assessing learning outcomes." 2020. [Online]. Available: https://champlain.instructure.com/courses/200147/pages/curriculum-mapping

Factors Influencing Computing Students' Readiness to Online Learning for Understanding Software Engineering Foundations in Saudi Arabia

Abdulaziz Alhubaishy
College of Computing and Informatics
Saudi Electronic University
Medinah, Saudi Arabia

Abstract—**The spread of Coronavirus disease (COVID-19) has enforced most universities/institutions over the world to transform their educational models (face-to-face and blended) bearing in mind the online educational environments as a temporary substitute. Consequently, all universities/institutions in Saudi Arabia have requested their students to continue the learning process using online environments. This transition has provided an opportunity to deeply investigate possible challenges as well as factors that influence the adoption of online learning as a future educational model for undergraduate students. This research measures the current undergraduate students' readiness for online learning and investigates factors that influence their level of readiness. Firstly, the research proposes the adoption of a validated multidimensional instrument to measure undergraduate students' readiness for online learning in different universities. Secondly, the research elaborates the findings by an in-depth study that highlights the main obstacles that hinder computing students' readiness to learn Software Engineering (SE) foundations using online learning. The research adopts survey research to measure students' readiness and analyzes the data to extract the readiness levels of different dimensions of the adopted instrument. Furthermore, interviews were conducted to specify the influential factors on computing students' readiness levels regarding learning SE foundations. Results show that students' readiness level for online learning is within the acceptable range while some improvements are needed. Furthermore, the study found that students' cognition, willingness, ignorance, and the amount of assistant and help they receive play a significant role in the success/failure of the adoption of learning SE foundations through online environment.**

Keywords—Readiness to online learning; software-engineering education; improving online learning; university students

I. Introduction

Online learning has brought many advantages to the discipline of education during this era. The continuous improvement of computers and the Internet has a positive impact on the growing adoption of online learning strategies by different educational institutions. The adoption of online learning is primarily related to different disciplines and dimensions, such as student's attitude and perception of tools [1]. Teaching and learning are also affected by cultural differences [2]. Different studies have drawn different results based on the investigated culture/region. Different scales have been developed to measure the influence of these dimensions on different cultures. Measuring the influence promotes new insights and exposes challenges that hinder the success of adopting online learning.

Software Engineering (SE) education is the area that requires using effective methods to ensure that theoretical foundations are linked to practical experience to narrow the gap between academia and industry [3]. The adoption of online education has motivated researchers to investigate how it can contribute to narrow this gap. Online SE education requires assuring the readiness of both educators and learners.

This research, in line with other researches in this area, seeks to investigate how undergraduate students are prepared to accept the online learning model for the future. It has adopted five dimensions measurement model, namely students' motivation, self-directed learning, control, computer/internet self-efficacy, and online communication self-efficacy, in order to explore the weak areas that might delay the success of adopting online learning models. To the best of our knowledge, all proposed studies are either outdated or have targeted participants from fewer number of institutions. Exploring these dimensions within the Saudi culture in different institutions provides us the current level of readiness along with insightful ideas on how to improve students' readiness for online learning and overcome possible challenges that might hinder learning some practical concepts such as SE foundations. Therefore, the various objectives of this research can be summarized in the following points:

1) Understanding the current abilities and attitudes of students.
2) Identifying students' motivation towards online learning.
3) Highlighting the influence of the non-linear sequence of studying materials using online environments as compared to a traditional linear sequence.
4) Understanding the current state of students' perception and ability to use technology.
5) Acquiring students' communication self-efficacy using online learning environments.
6) Investigating the challenges and factors that impact undergraduate students' readiness for learning SE foundations through online methods and how to overcome these challenges.

II. Literature Review

A. Online Learning Readiness

Readiness for online learning is a concept that is mainly defined as the learner's confidence in online communication, preferences of the online model over the traditional model, and the ability to be involved in online learning [4]. Researchers have validated different scales to measure student's readiness. One of the most adopted scales by the literature is the five dimensions scale developed by [5]. These measured dimensions include students' self-directed learning, motivation for learning, control, computer and internet self-efficacy, and online communication self-efficacy.

Each dimension in the scale has been extensively elaborated and defined by many authors, while a separate scale for each of these dimensions has been proposed and validated. For example, self-directed learning is the concept that encompasses students' initiative to discover their needs and learning goals besides identifying resources for learning and choosing the best method to implement strategies of learning and eventually evaluate their outcomes [4-6].

Sethi et al. have used a pre-validated scale for measuring the readiness of 789 students from three different universities in Pakistan [6]. The authors have concluded that there is a need for training female students about handling computers and the Internet. Furthermore, improving learner's motivation and online interaction are two issues that course designers must take into consideration.

In 2011, Chanchary and Islam have studied the readiness of three groups at Saudi university [7]. The study has found that despite the ability of students to use applications and tools, they were not comfortable in online communications. Interestingly, 73% students of the studied groups have expressed their unwillingness to adopt online learning as a substitute for traditional learning.

Recently, readiness of a total of 204 female students was measured at Princess Nourah Bint Abdulrahman University (PNU) in Saudi Arabia[8]. The study has found that participants had a high level of readiness towards online learning as a substitute for the traditional closed-circuit distance technology.

B. Online-Learn Software Engineering

The main issue in teaching SE fundamental courses is teaching theoretical foundations and relating them to a hand-on experience to narrow the gap between academia and industry [3]. Most universities tend to dedicate most of the time and effort during the semester for teaching theoretical part of the course. Researchers and practitioners work on narrowing the gap by creating a balance between theoretical and practical experience. However, this is a hard task because of many challenges and difficulties that are encountered by teachers and students per se.

Heredia et al. have surveyed the literature to elicit the current pedagogical teaching approaches used in teaching software process [9]. The authors have found that most of methods are lectures with few number of exercises or projects or both. The other approaches include gamification as a type of game-related methods, realism, simulation, and missing subjects. In addition to these approaches, Marques et al. have highlighted six categories of pedagogical approaches for teaching practical SE; case studies, open source, learn by doing, problem-based learning, service learning, and inverted classroom [10].

Despite the benefits of these pedagogical approaches, some challenges and obstacles still exist. For example, game-related methods approach has the ability to enhance students' motivation and their learning outcome [11]. However, it has created new challenges, such as the cost of the approach and the lack of empirical data, to examine its benefits along with the lack of unified steps for educators in SE [12]. Based on a recent systematic review, Garousi et al. have categorized the obstacles during adopting different methods of teaching software testing into challenges related to instructors and students as well as the challenges related to instructors [13]. Within the first category, the authors have identified three challenges, which are low motivation, learning new tools, and the need to increase cognitive load of students. Within the later category, course design has been identified as the main challenge in addition to time constraints, resources constraints, assessing students' progress, and the challenge of integrating testing with other courses.

Online teaching of SE adds the aforementioned challenges into the main challenges being faced while adopting pedagogical online teaching approach. Though online teaching pave the way for SE educators to enhance the utilization of new teaching methods over traditional ones that are time consuming and do not allow instructors to fully teach concepts during the class time, such as the tools for software testing education [13]. For example, it has been found that most of the courses that are taught using non-traditional pedagogical approaches, such as online and blended courses, and adopted the game-related approach in SE education have utilized the approach completely during the online part [12]. Among different areas in SE, software design, software process, requirement engineering, and testing were highlighted in the literature as the most taught areas using online approach [14].

By considering which methods can be utilized during online teaching of SE, there is a need to highlight the obstacles being faced while online teaching and synthesize them with obstacles of teaching SE in order to have a full understanding of the challenges that SE online education faces. For example, the level of computer and internet self-efficacy as a dimension for measuring students' readiness to online learning can influence the successful adoption of SE tools during online courses. This can be true when a SE course is taught during early semesters of an undergraduate program where level of computer and internet self-efficacy of some students is low which eventually hinders the success of utilizing SE methods through online courses. Another example is the learners' experience and readiness to learn by considering the age of learners; adult or teenager, and how these factors can influence the learning process [15]. Mora et al. have proposed and tested a gamification framework on Requirement Engineering (RE) course taught completely online for adult students (age $>$ 25) [16]. The researchers concluded that the proposed framework was able to improve the level of students' engagements and motivation during the online course.

III. RESEARCH METHOD

To achieve the research objectives, we adopted a mixed-method research called explanatory sequential design strategy, which means that we adopted quantitative research followed by qualitative research as stated by [17]. Adopting the explanatory sequential design strategy allows us to highlight the weak areas that hinder undergraduate students' readiness to online learning besides understanding why and how these weak areas actually hinder the success of adopting the online learning for learning SE. Within a single study, Creswell have explained that adopting explanatory mixed method necessitates building the qualitative study based on the results of data analyzed via the quantitative phase [18]. The advantage of adopting this method lies in the fact that the research problem needs "more in-depth understanding of the quantitative results (often cultural relevance)" [18].

The first phase is accomplished by adopting the quantitative research in the form of survey strategy which concerns with validating a measurement model for students' readiness by controlling the activities of designing and disseminating questionnaires. The collected data measures the five dimensions of learners' readiness for online learning as proposed by [1]. In addition, demographic information from participants is collected to analyze the potential influence on their responses.

Unlike the number of studies conducted by different authors, such as [7] and [8], in which they have examined online readiness within a fewer number of regions, we examined online readiness of students by considering the five main regions in Saudi Arabia; Central region, Western region, Eastern Region, Southern region, and Northern region. The sample for this quantitative survey allows us to enhance our understanding of the current level of online readiness over a broader region.

We then adopted the qualitative study in the form of semi-structured interviews to understand how the extracted weak dimensions have influence on students' readiness for online learning, especially for learning SE concepts. The interview allows us to gather information in different formats than when a quantitative strategy was used. Targeted information encompasses the challenges and reasons behind the weak dimensions and how they hinder the application of learning SE courses. Therefore, semi-structured interviews are adopted where a set of open-ended questions is designed based on the quantitative results as the explanatory sequential design suggests.

The interview questions have been created and divided into two main sections. The first section concerns with collecting information regarding the previous (i.e. before COVID-19 Outbreak) uses of any educational environment, online tool, or software to help students understand SE concepts. The section also extracts information about current (during COVID-19 Outbreak) uses of these online environments and tools. The main reason of this section is to understand whether there were early adoptions of online tools to teach SE concepts, or there was a substantial differentiation between the uses of these tools before and after the pandemic. The second section concerns with collecting more information on how weak levels of online readiness dimensions influence the utilization of teaching SE concepts and how improving all dimensions level leads to a better utilization. Analyzing the interview takes place through thematic analysis to extract the main themes that influence the level of online readiness. The thematic analysis process was followed as illustrated by [19], and the main themes were extracted as explained in Section IV-B.

IV. Results

A. Results of Quantitative Study

Over a period of three months, a total of 244 valid responses were collected and analyzed. Collected data shows diversity in some demographic variables. Regarding the gender, around 82% (N=200) of the responses were male and 18% (N=44) were female. In Saudi universities, female and male students are studying at different campuses. Even though the survey was conducted through online means, reaching to a larger pool of female students was obstacle because of the separation strategy in studying and teaching. Undergraduate participants constituted the majority of responses 97.13% (N= 237), and only 2.87% (N= 7) were graduate students. This is because of the small number of graduate students as compared to undergraduate students. Regarding the year of study, around 60% of participants were in their first two years of study. Finally, even though the equal effort was paid to gather responses from all regions, respondents from Eastern, Western, and Central universities constituted more than 80% of total respondents. The least number of respondents were received from Northern universities followed by Southern universities. However, number of students in the later regions were much less than other three regions. Table I shows the demographic information of the survey respondents.

Reliability testing of the survey constructs was carried out by examining Cronbach's Alpha. A reliable construct shows Cronbach's Alpha value of 0.70 or above as stated by [20]. All five constructs showed acceptable values ranging from 0.70 to 0.79 as depicted in Table II. Learner Control reached the lowest acceptable value (0.70) followed by computer/Internet self-efficacy (0.74). Relating this level of reliability to the original scale, we found that these two constructs suffered from the convergent validity where the values of AVE did not reach the threshold value of 0.50 as suggested by [21].

Table III shows the mean of each item along with the overall mean for each construct. All individual items scored above acceptable means (above 3.5), except one item "SDL3: I manage time well" which scored the lowest mean of 3.35. Overall means of constructs were found to be 3.51, 3.72, 3.72, 3.94, and 4.09 for Learner control, Self-directed learning, Online communication self-efficacy, Computer/Internet self-efficacy, and Motivation for learning respectively. Based on the five-point Likert scale we used in the survey, we divided the responses into four intervals. We then compared the mean of each construct with a proposed model that was introduced by [22]. However, the proposed model by [22] examines readiness by measuring different dimensions, which are Technology, Innovation, People, and Self-Development. Means falling within the first interval ($\bar{X}$: 1.0 - 2.6) indicate that the students are not ready for the e-learning yet and need a lot of work regarding the respected dimension reflecting that mean. Means falling within the second interval ($\bar{X}$: 2.6 - 3.4) show that the dimension needs some work before becoming ready for online learning. Means that fall within the third interval ($\bar{X}$: 3.4 - 4.2) and fourth interval ($\bar{X}$: 4.2 - 5.0) indicate the readiness, but with the need for some improvements to dimensions falling within the third interval. Therefore, the threshold of $\bar{X}$ should be more than or equal to 3.4. Fig. 1 shows the means of our survey constructs illustrating that all constructs achieved the threshold of accepted level of online readiness. The mean of

TABLE I. DEMOGRAPHIC VARIABLES

Variable	Category	Frequency	Percentage (%)
Gender	Male	200	82
	Female	44	18
Education Level	Undergraduate	237	97.13
	Graduate	7	2.87
Year of Study	Preparatory/First Year	37	15.1
	Second	117	48.0
	Third	30	12.3
	Forth	21	8.6
	Fifth	39	16.0
University Location (Region)	Western	84	34.4
	Eastern	89	36.5
	Central	39	16
	Southern	17	7
	Northern	15	6.1

TABLE II. CRONBACH'S ALPHA RELIABILITY FOR THE SURVEY CONSTRUCTS

Construct	Number of Items	Cronbach's Alpha
Computer/Internet self-efficacy	3	0.74
Self-directed learning	5	0.79
Learner control	3	0.70
Motivation for learning	4	0.75
Online communication self-efficacy	3	0.79

each contracts ranges from $\bar{X} = 3.51$ to $\bar{X} = 4.09$, which denotes that all our constructs fall within the third interval where some improvements are needed.

B. Results of Qualitative Study

All survey participants were recruited to participate in the interviews. Of the pool of the participants, nine participants showed their interest to participate in the qualitative study, so the interviews were arranged and conducted. After a comprehensive analysis of the participants' interviews using the thematic analysis, four main themes were extracted and considered as the main barriers to the successful adoption of online SE learning methods. These themes are:

1) Students' cognition towards new learning methods.
The overall perspective of many students is that the online learning cannot fully help students understand different SE concepts; rather, traditional learning is the most suitable method. Some of the codes that were extracted from participants regarding their cognition towards new learning methods include resistance to change, fear of technology, and lack of motivation for searching and practicing new online learning tools. Firstly, fear of transition from traditional learning methods, i.e. face-to-face learning, to a technology-based method appears as one of the most influential factors. Students are also suffering from the lack of encouragement and motivation to learn and search for new tools that help them understand different SE concepts. These factors have been identified as assistant contributors to adherence tendency towards traditional methods.

2) Student's ignorance of the tools and their benefits.
A main highlighted theme is that students are actually not aware of the availability of different online teaching methods and tools and their benefits, and are unable to practice them because of lack of knowledge. It has been identified that almost most of online tools and methods, such as game-related methods, are not recognized by students. Still, theoretical explanations using traditional methods represent the main teaching style in many universities. Furthermore, students are actually not aware of different benefits that the online tools and methods can provide them. Within this theme, the awareness about the availability and benefits of these methods and tools has been recognized by a minor number of students, however it has been highlighted that their knowledge along with other contributing factors, such as lack of guidance, can prevent them from adopting these methods and tools at their own.

3) Students' willingness to follow traditional learning methods.
An interesting theme constructed from different codes concerns about the willingness of students to learn traditionally. The reasons behind this are based on their current experience of online learning where only face-to-face lectures are transferred to virtual classes during COVID-19 pandemic, and they have the lack of willingness to online learning, inability to link the learning outcome with online tools, lack of online communication and social interaction, and believes in lack of knowledge retention. Each of these reasons might hinder the successful adoption of any learning strategy that is based on online tools and methods.

4) Instructors' and departments' tendency towards new teaching methods and tools.
An important theme is related to instructors' behaviour towards teaching their students. Different students have emphasized that their instructors are following the traditional methods, and they have no tendency to provide them any resources or information rather than verbally explaining the concepts. Therefore, regarding the instructors' attitude, it has been found that most of them tend to traditional methods. Furthermore, there is the lack of provided resources by the department/college and inability to accomplish online learning activities because of lack of support and assistance.

V. DISCUSSION

An intrinsic relationship was highlighted between these themes and the level of online readiness. Firstly, even though

TABLE III. Mean scores for the online readiness to learn scale

Construct	Item	$\bar{X}$	SD
Computer/Internet self-efficacy	CIS1: I feel confident in performing the basic functions of Microsoft Office programs (MS Word, MS Excel, and MS PowerPoint).	3.77	1.13
	CIS2: I feel confident in my knowledge and skills of how to manage software for online learning.	3.89	1.11
	CIS3: I feel confident in using the Internet (Google, Yahoo) to find or gather information for online learning.	4.16	1.09
	Overall for CIS:	3.94	1.11
Self-directed learning	SDL1: I carry out my own study plan.	3.96	1.12
	SDL2: I seek assistance when facing learning problems.	3.85	1.18
	SDL3: I manage time well.	3.35	1.21
	SDL4: I set up my learning goals	3.52	1.19
	SDL5: I have higher expectations for my learning performance.	3.93	1.11
	Overall for SDL:	3.72	1.16
Learner control	LC1: I can direct my own learning progress.	3.79	1.08
	LC2: I am not distracted by other online activities when learning online (instant messages, Internet surfing).	3.21	1.35
	LC3: I repeated the online instructional materials on the basis of my needs.	3.54	1.16
	Overall for LC:	3.51	1.20
Motivation for learning	MFL1: I am open to new ideas.	4.13	1.03
	MFL2: I have motivation to learn.	4.30	1.02
	MFL3: I improve from my mistakes.	4.20	1.00
	MFL4: I like to share my ideas with others.	3.73	1.25
	Overall for MFL:	4.09	1.08
Online communication self-efficacy	OCS1: I feel confident in using online tools (email, discussion) to effectively communicate with others.	3.90	1.21
	OCS2: I feel confident in expressing myself (emotions and humor) through text.	3.57	1.30
	OCS3: I feel confident in posting questions in online discussions.	3.69	1.26
	Overall for OCS:	3.72	1.26
	Overall	3.82	1.56

students' level of confidence in their knowledge and skills of how to manage a software for online learning is acceptable ($\bar{X}$: 3.89) with the need for improvement, it has been found that students lack this level of confidence when it comes to a specific online tool, such as SE online software and tools. The reasons are mainly related to their cognition and/or ignorance of the new tools and methods. Therefore, learning SE concepts through online tools and methods is mainly affected by their confidence in skills and knowledge along with their confidence in knowledge and skills related to the designated tools for learning SE concepts.

Self-directed learning of undergraduate students was also found to be acceptable ($\bar{X}$: 3.72), but it needs some improvements. This has been found in the qualitative study related to the level of students' willingness to proceed with learning SE concepts using new tools and online learning methods. For example, managing the time is a main issue that was found where many students are not able to manage their time during learning and experiencing new online tools. The improvement can be ensured by providing students required training and workshops to enable them to effectively manage their time during online learning. Furthermore, students' willingness to seek for assistant when facing problems is lower than expectation. Therefore, it can be another area of improvement that can be handled.

Learner control is the weaker dimension that has been highlighted ($\bar{X}$: 3.51) which needs more improvements to enhance the readiness of undergraduate students. Learner control has been mainly identified with two themes; students' cognition and willingness. Within the first, students are mainly dependent on their experience of traditional learning progress and are not familiar with the management technique to direct their learning progress when it comes to online learning of SE tools. Furthermore, students are mainly dependent on repeated instructions provided by instructors rather online instructional and guidance procedures on how to understand and practice the online tools. Within the later, students' willingness to online learning is claimed to be low because of the distraction and inability to focus during online learning. Distraction by other activities is considered the second most influential factor ($\bar{X}$: 3.21) that needs to be considered and improved by adopting corrective action to attract students towards avoiding distraction during online learning. The correction actions can be focusing more on online course design, choosing the right tools, and motivating students.

Motivation for learning was the highest dimension ($\bar{X}$: 4.09). Interestingly, when it comes to online learning, the qualitative study exhibits a low motivation among undergraduate students. Different students have explained that they are not as motivated to online study and using new tools as traditional learning. For example, a student has explained that his university has transferred the lab session virtually and used the repl.it software which is a platform for creating and sharing code online [23]. The student has mentioned "Even though we were encouraged to share our codes with the instructor, a few numbers of us have done that". The student continued explaining the situation by saying "We still do not feel that the instructor's online correction of our code motivates us the same amount as it has been done during the face-to-face corrections".

Finally, online communication self-efficacy was acceptable ($\bar{X}$: 3.82) with the need for some improvements. The qualitative study has highlighted that the level of confidence by using online tools for communication is high, but the issue arises

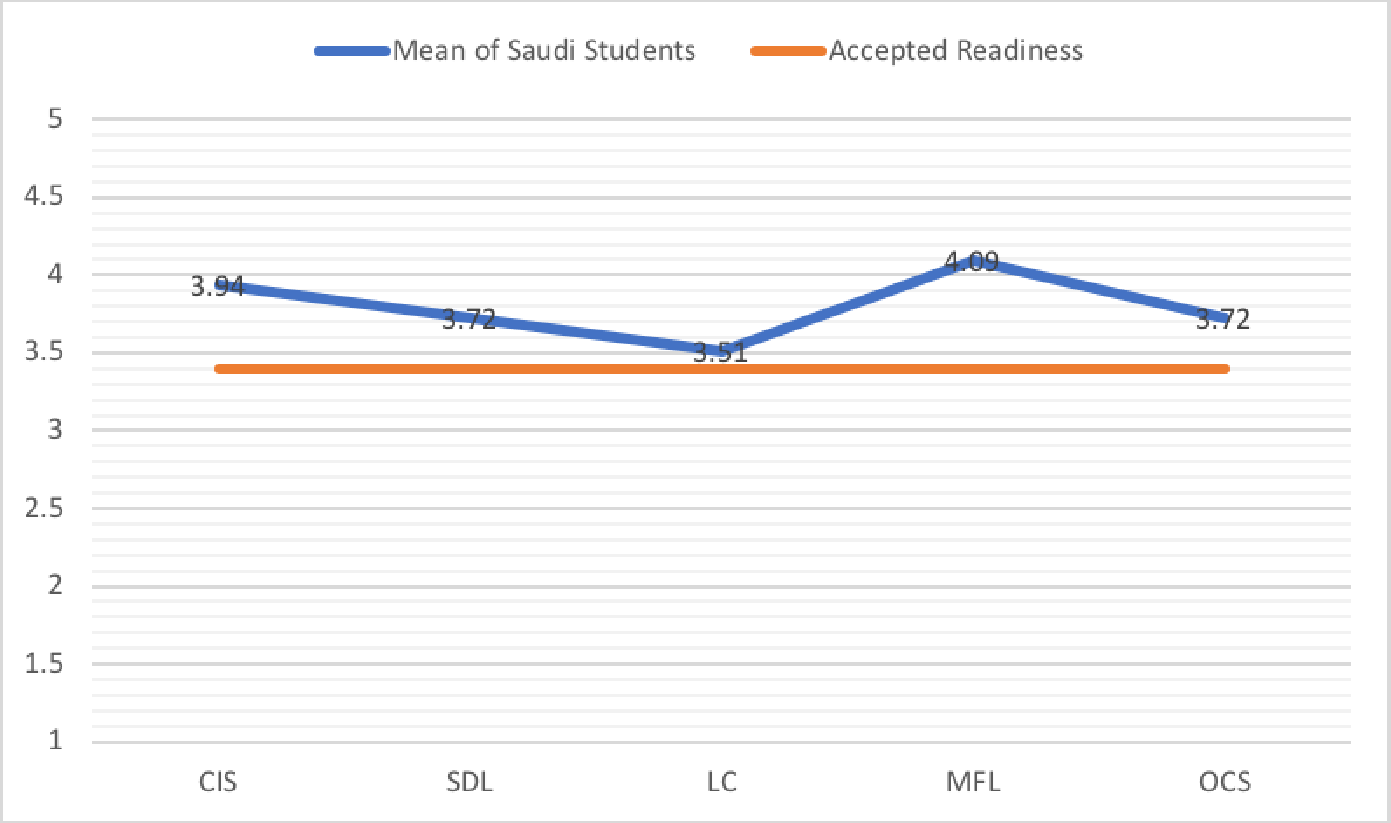


Fig. 1. Current Students' Readiness Level Compared to the Accepted Level

when students experience a problem with learning new tools. For instance, it has been found that the level of confidence in posting questions using online discussions is lower because of two reasons. Firstly, students avoid questioning because they are not sure about getting the right and timely response they seek for. Secondly, students might have the fear that the answers to their questions are already provided elsewhere and they should look for the answers. In both cases, motivating students and increasing the level of their willingness and courage towards online learning along with providing them necessary assistant and support by instructors, technical support team, and department can lead to improve students' online communication self-efficacy.

VI. Conclusion

The improvements that have been adopted in education system in Saudi Arabia had a significant impact on the development of undergraduate students' skills and knowledge. Readiness to online learning is one of the most developed areas, where the level of students has been measured in this study. However, the need for more improvements is required where increasing this level of readiness can lead to better learning and understanding of practical concepts, such as SE concepts, during the adoption of online learning methods. This study contributed by measuring the current level of undergraduate students throughout the adoption of one of the most recognized readiness scales. Furthermore, the study contributed by investigating different dimensions of readiness and methods for improvements besides exploring how different factors correlate to constitute barriers that hinder the improvement in the readiness level among undergraduate students.

The recommendations of this research are multifaceted. Firstly, decision makers in educational institutes must adhere to the set of practices and processes that enable students to utilize online learning tools and methods and persevering high motivation. The practices can entail having well-trained instructors for online teaching, designing curriculum that links learning outcome with the new online tools and methods, and working on continuous support and help for students during the online learning process. Secondly, enriching students with the required knowledge and skills to overcome different challenges that they might face during the online learning. Students also need training courses and workshops to help them manage their time wisely and effectively during the online learning. Finally, faculty members are also required to have the necessary training and obtain the required knowledge and skills regarding online teaching to assist their students effectively.

Acknowledgment

The author extends his appreciation to the Deanship of Scientific Research at Saudi Electronic University for funding this work.

References

[1] R. G. Saadé, X. He, and D. Kira, "Exploring dimensions to online learning," *Computers in human behavior*, vol. 23, no. 4, pp. 1721–1739, 2007.

[2] J. Küsel, F. Martin, and S. Markic, "University students' readiness for using digital media and online learning—comparison between germany and the usa," *Education Sciences*, vol. 10, no. 11, p. 313, 2020.

[3] A. Alsolamy and R. Qureshi, "The proposal of a qualification based approach to teach software engineering course." *International Arab Journal of Information Technology (IAJIT)*, vol. 12, no. 2, 2015.

[4] D. Warner, G. Christie, and S. Choy, "Readiness of vet clients for flexible delivery including on-line learning," *Brisbane: Australian National Training Authority*, 1998.

[5] M.-L. Hung, C. Chou, C.-H. Chen, and Z.-Y. Own, "Learner readiness for online learning: Scale development and student perceptions," *Computers & Education*, vol. 55, no. 3, pp. 1080–1090, 2010.

[6] A. Sethi, A. Wajid, and A. Khan, "E-learning: Are we there yet?" *Professional Medical Journal*, vol. 26, no. 4, 2019.

[7] F. H. Chanchary and S. Islam, "Is saudi arabia ready for e-learning? a case study," in *The 12th International Arab Conference on Information Technology, Naif Arab University for Security Sciences, Saudi Arabia, retrivied: January*, vol. 20, 2011, p. 2017.

[8] A. Alahmari and R. J. Amirault, "The use of e-learning in highly domain-specific settings: Perceptions of female students and faculty in saudi arabia," *Quarterly Review of Distance Education*, vol. 18, no. 4, pp. 37–56, 2017.

[9] A. Heredia, R. C. Palacios, and A. de Amescua Seco, "A systematic mapping study on software process education." in *SPETP@ SPICE*, 2015, pp. 7–17.

[10] M. R. Marques, A. Quispe, and S. F. Ochoa, "A systematic mapping study on practical approaches to teaching software engineering," in *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*. IEEE, 2014, pp. 1–8.

[11] D. Dicheva, C. Dichev, G. Agre, and G. Angelova, "Gamification in education: A systematic mapping study." *Journal of Educational Technology & Society*, vol. 18, no. 3, 2015.

[12] M. M. Alhammad and A. M. Moreno, "Gamification in software engineering education: A systematic mapping," *Journal of Systems and Software*, vol. 141, pp. 131–150, 2018.

[13] V. Garousi, A. Rainer, P. Lauvås jr, and A. Arcuri, "Software-testing education: A systematic literature mapping," *Journal of Systems and Software*, p. 110570, 2020.

[14] K. Wendt, "Audience and content areas of online software engineering education and training: A systematic review," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[15] T. Carney, "Andragogy in action: Applying modern principles of adult learning," *Canadian Journal of Communication*, vol. 12, no. 1, 1986.

[16] A. Mora, E. Planas, and J. Arnedo-Moreno, "Designing game-like activities to engage adult learners in higher education," in *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality*, 2016, pp. 755–762.

[17] R. B. Johnson and A. J. Onwuegbuzie, "Mixed methods research: A research paradigm whose time has come," *Educational researcher*, vol. 33, no. 7, pp. 14–26, 2004.

[18] J. W. Creswell and J. D. Creswell, *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2017.

[19] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative research in psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[20] J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, R. L. Tatham *et al.*, *Multivariate data analysis*. Prentice hall Upper Saddle River, NJ, 1998, vol. 5, no. 3.

[21] C. Fornell and D. F. Larcker, "Structural equation models with unobservable variables and measurement error: Algebra and statistics," 1981.

[22] C. H. Aydın and D. Tasci, "Measuring readiness for e-learning: Reflections from an emerging country," *Journal of Educational Technology & Society*, vol. 8, no. 4, pp. 244–257, 2005.

[23] T. Tang, S. Rixner, and J. Warren, "An environment for learning interactive programming," in *Proceedings of the 45th ACM technical symposium on Computer science education*, 2014, pp. 671–676.

Road Traffic Accidents Injury Data Analytics

Mohamed K Nour[1]
College of Computer and
Information Systems
Umm Al-Qura University

Atif Naseer[2]
Science and Technology Unit
Umm Al-Qura University

Basem Alkazemi[3]
College of Computer and
Information Systems
Umm Al-Qura University

Muhammad Abid Jamil[4]
College of Computer and
Information Systems
Umm Al-Qura University

***Abstract*—Road safety researchers working on road accident data have witnessed success in road traffic accidents analysis through the application data analytic techniques, though, little progress was made into the prediction of road injury. This paper applies advanced data analytics methods to predict injury severity levels and evaluates their performance. The study uses predictive modelling techniques to identify risk and key factors that contributes to accident severity. The study uses publicly available data from UK department of transport that covers the period from 2005 to 2019. The paper presents an approach which is general enough so that can be applied to different data sets from other countries. The results identified that tree based techniques such as XGBoost outperform regression based ones, such as ANN. In addition to the paper, identifies interesting relationships and acknowledged issues related to quality of data.**

Keywords*—*Traffic Accidents Analytics (RTA); data mining; machine learning; XGBOOST

I. Introduction

Road Traffic Accident (RTA) is an unexpected event that unintentionally occurs on the road which involves vehicle and/or other road users that causes casualty or loss of property. Over 90% the world's fatalities on roads occur in low and middle income countries which account for only 48% of world's registered vehicles [1]. The financial loss, which is about US$518 billion, is more than the development assistance allocated for these countries. While developed rich nations have stable or declining road traffic death rates through coordinated correcting efforts from various sectors, developing countries are still losing 1–3% of their gross national product (GNP) due to the endemic of traffic casualties. World Health Organization (WHO) fears, unless immediate action is taken, road crash will rise to the fifth leading cause of death by 2030, resulting in an estimated 2.4 million fatalities per year [1].

Thus, measures to reduce crashes based on in-depth understanding of the underlying causes are of great interest for researchers.The 21st century has been seeing a rapid growth of road motorisation due to rapid increase of population, massive urbanisation, and increased mobility of the modern society, risks of road traffic fatality (RTF) may also become higher and RTA can also be assumed as a “modern epidemic”. This paper presents an analytic framework to predict accident severity for road traffic accidents [1]. Past research on road traffic accidents analysis had mainly relied on statistical methods such as linear and Poisson regression. This paper presents an analytic framework to predict accident severity for road traffic accidents. In particular, the paper addresses issues related to data preprocessing and preparation such as data aggregation, transformation, feature engineering and imbalanced data. In addition, the paper aims to apply machine learning models to enable more accurate predictions. Hence, the compares the performance of several machine learning algorithms in predicting the accident injury severity. In particular, the paper applies logistic regression, support vector machines, decision trees, random forest , XGBoost and artificial neural network models. The rest of this paper is organized as follows: Section II introduces some previous works. Section III shows the methodology used in this work, Section IV describes the data management and the patterns of traffic accident data. Section V shows the results and analysis of the all the approached used in this work. Section VI gives the conclusions and future works.

II. Literature Review

Mehdizadeh et al. [2] presented a comprehensive review on data analytic methods in road safety. Analytics models can be grouped into two categories: predictive or explanatory models that attempt to understand and quantify crash risk and (b) optimization techniques that focus on minimizing crash risk through route/path-selection and rest-break scheduling. Their work presented a publicly available data sources and descriptive analytic techniques (data summarization, visualization, and dimension reduction) that can be used to achieve safer-routing and provide code to facilitate data collection/exploration by practitioners/researchers. The paper also reviewed the statistical and machine learning models used for crash risk modelling. Hu et al. [3] categorized the optimization and prescriptive analytic models that focus on minimizing crash risk. Ziakopoulos et al. [4] critically reviewed the existing literature on different spatial approaches that include dimension of space in its various aspects in their analyses for road safety. Moosavi et al. [5] identified weaknesses with road traffic accidents research which include: small-scale datasets, dependency on extensive set of data, and being not applicable for real time purposes. The work proposed a data collection technique with a deep- neural-network model called Deep Accident Prediction(DAP); The results showed significant improvements to predict rare accident events. Zagorodnikh et al.[6] developed an information system that displays the accidents concentration on electronic terrain map automatically mode for Russian RTA to help simplifying the RTA analysis.

Kononen et al. [7] analysed the severity of accidents occurred in United States using logistic regression model. They reported performance 40% and 98%, for sensitivity and specificity respectively. Also, they identified the most important predictors for injury level are: change in velocity, seat belt use, and crash direction.

The Artificial Neural Networks (ANNs) are one of the data mining tools and non-parametric techniques in which researchers have analysed the severity of accidents and injuries among those involved in such crashes. Delen et al. [8] applied a ANNs to model the relationships between injury severity levels and crash related factors. They used US crash data with 16 attributes. The work identified four factors that influence the injury level: seat belt, alcohol or drug use, age, gender, and vehicle.

Naseer et al. [9] introduces a deep learning based traffic accident analysis method. They highlighted deep learning techniques to build prediction and classification models from the road accident data.

Sharma et al. [10] applied support vector machines with different Gaussian kernel functions for crash to extract important features related to accident occurrence. The paper compared neural network with support vector machines. The paper reported that SVMs are superior on accuracy. However, the SVMs method has the same disadvantages of ANN in traffic accident severity predication as mentioned earlier

Meng et al. [11] used XGBoost to predict accidents using road traffic accident data from multiple sources. They used historical data along with weather and traffic data. Schlogl et al. [12] performed multiple experiments to proof that XGBoost performs better as compared to several machine learning algorithms.

Ma et al. [13] proposed the XGBoost based framework which analysed the relationship between collision, time and environmental and spatial factors and fatality rate. Results show that the proposed method has the best modelling performance compared with other machine learning algorithms. The paper identified eight factors that have impact on traffic fatality.

Cuenca et al. [14] compared the performance of Naive Bayes, Deep Learning and Gradient Boosting to predict the severity of injury for Spanish road accidents. Their work reported that Deep Learning outperform other methods

III. Methodology

The methodology adopted in this paper is shown in Fig. 1. The first step for data analytics process is data collection which is regarded as the primary building block for successful data analysis project. There are many data sources like sensors, visual data through cameras, and IoT and mobile devices which captures data in different formats and need to be stored realtime or offline. In addition, data collected from different authorities related to traffic volume, accident details and demographic information. The storage can be on the local servers or on cloud. The key of data management pyramid is data preprocessing. The data acquired from the storage locations cannot be used as it is. It requires preprocessing before performing any analysis. The acquired data may include missing information that needed to rectify as well as many information needed to be removed due to duplication. The preprocessing may involve the data transformation as it helps in data normalization, attribute selection, discretization, and hierarchy generation. Data reduction maybe required on the large scale of data as the analysis of a huge amount of data is harder. Data reduction increases the efficiency of storage and the analysis cost. Data Analysis use multiple machine learning algorithms to get the insight of data. The data analysis is very crucial for any organization as it provides the detailed information about the data and is helpful in certain decision making and predictions about the business. Data can be presented in various forms depending on the type of data being used. The data can be shown into organized tables, charts, or graphs. Data presentation is very important for business users as it provides the results from the analysis of data in a visual format.

One of the most important tasks for road risk analysis and modelling is to predict accident severity level. This paper looks at building predictive model for accident severity level and investigate the process of constructing a classification model to predict accident severity level, in particular the study:

- Presents the data management framework. This is followed by discussion on how the data was prepared prior to modelling. This includes pre-processing and data cleansing. This section is presented in the Data Management Framework section
- Identifies gaps in the RTAs predictive modelling techniques. This section gives brief background on each technique used in this paper and presents prior work in road traffic accident prediction together with with data requirements and recorded performance results. This topic is presented in the Data Analysis section
- Build prediction models. This section begins by stating performance metrics then followed by data used. Then compares classifiers in particular; logistic regression, support vector machines, neural networks, decision trees, random forest and Extreme gradient boosting tree (XGboost) . In this section, the unbalanced class distribution is investigated to see their impact on injury severity during accidents. This is presented in the results section

IV. Data Management

A. Data Collection

The data comprises publicly available data from UK government which spans the period of 2005 to 2019 [15]. Although the UK department of transport provide data from 1979, it was reported the data collected from 2005 onwards are more accurate and contains less missing data. The records shows information on road traffic collisions that involve personal injury occurring on public roads which have been reported to the police. Data is collected by the authorities at scene of an accident or, in some cases, reported by a member of the public at a police station, then processed and passed on to the authorities. Data includes, 2 million unique collisions, with x, y space coordinates available. Data related to traffic flow is and information about all UK network roads and local authorities are also available separately. The dataset contains a single entry for each accident with 33 attributes (features). The attributes can be grouped by geography, accident-focused, weather, time . Data Related to vehicles involved with the accidents is stored in separate file with 16 attributes. Data related to casualty involved with the accident is stored in 23 fields file. The relation between these three files is one to many, i.e one accident row can contain many casualty rows and many vehicle rows with the accident index is the linking field.

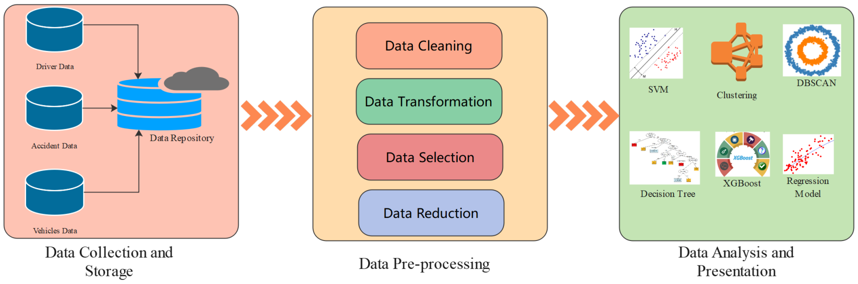


Fig. 1. Proposed Methodology

The severity in any accident is the most important feature to analyze the injury pattern. According to the WSDOT [16], the severity level during the accidents are measured using the KABCO scale, which uses the parameters: fatal (K), incapacitating-injury (A), non-incapacitating injury (B), minor injury (C), and property damage only (PDO or O)).

In this work, we divided the accidents into three categories i.e. Fatal (where the death occurs within 30 days of accident), Serious injury (where the person requires hospital treatment) and Slight injury (where the person not required any medical treatment).

In this project the data is stored in relation database. In future work however, the relational data first will be denormalized then transformed into Hadoop key-value records

B. Data Preprocessing

Different data preprocessing and cleaning methods were applied to the data. Data pre preprocessing involve many tasks and techniques. This include: dealing with missing value, outlier or anomaly detection, feature selection. A key stage in the data analytic is the selection of data. Data needs to be of good quality and clean.

Data quality considerations include accuracy, completeness and consistency [17]. In addition data volume is important as well. Data should be large enough to be of value in predictive modelling. It must be split into training, test and validation subset in order to evaluate the model. The following data preprocessing steps were applied to the data in order to make the data ready for analysis and machine learning algorithms:

- Most machine learning methods require the data to be in either binary or numeric format. However, in real life data sources include category attributes such road type, casualty class. All category attributes will be converted into numeric values.
- Numeric values differ in ranges. To avoid bias to large numeric values all numeric values will be normalized to values between 0 and 1
- Records with missing values will be removed.
- Date field will create several relate fields such as month, year and week.
- Determine less quality attributes. Attributes with more than 70% missing values will be removed.
- Calculate the correlation between severity level and all other attributes. Attributes with high correlation values will be removed as well as attributes with very low correlation values.
- Create fields related to easting and northing to create zones for accidents instead of specific location. The threshold value for zone level is $1km^2$.

One of the issues that faces building analytic models for crash severity, is the imbalance of data [18] where the occurrence fatality which is infrequent or rare event compared to no or minor injury accidents. Due to the extreme imbalance of accident data most algorithms will not produce good predictive models and perform poorly will likely miss-classify the fatal accidents as it is not prevalent in the dataset [19]. For imbalanced data sets such as traffic accidents data, sampling techniques can help improve classifier accuracy [19]. Two sampling techniques; undersampling and oversampling techniques will be discussed below.

Under sampling is used to adjust the class distribution of a dataset in favour of the minority class. With undersampling, the majority class is reduced or under sampled [17] and randomly eliminates data from the majority class until both classes match. Oversampling is a technique used in data mining to adjust the class distribution of a dataset in favour of the majority class [18]. Oversampling, on the other hand, the minority class increased or over sampled until the size meets that of the majority class. However, these techniques require specialised skill and it can take a significant time-frame to identify the best sample.

In this study, we need to apply feature selection task on the dataset. The dataset is preprocessed as specified in Tables I, II and III. The Table I shows the features with respect to accidents, Table II shows all features of vehicles, while the Table III highlights the features with respect to casualty with their type. The tables also shows the preprocessing on the features so that some features excludes from the list while some of them adjusted with scale.

TABLE I. ACCIDENTS FEATURES

Variable Name	Type	Preprocessing
Accident Index	Link field	EXECLUDE (unique in accidents)
Police Force	Number from 1-98	0 london 1 otherwise
Accident Severity	1 Fatal 2 Serious 3 Slight	EXECLUDE
Number of Vehicles	Numeric	SCALE
Number of Casualties	Numeric	SCALE
Date (DD/MM/YYYY)	DATE	Split into Moth, Year and Week, weekend Weekday
Day of Week	1 TO 7	SCALE
Time (HH:MM)	TIME	Split into Rush hours and Non Rush Hours
Location Easting OSGR (Null if not known)	Numeric	Remove Last two dists and scale
Location Northing OSGR (Null if not known)	Numeric	Remove Last two dists and scale
Longitude (Null if not known)	Numeric	INCLUDE
Latitude (Null if not known)	Numeric	INCLUDE
Local Authority (District)	1 to 941	exclude
Local Authority (Highway Authority - ONS code)	208 Items	exclude
1st Road Class	1 to 6	0 motoryway, 1 othewise
1st Road Number	Numeric	**exclude**
Road Type	1 to 12	1 motoryway
Speed limit	Numeric	SCALE
Junction Detail	0 TO 9	0 no juntion, 1 otherwise
Junction Control	0 TO 4	0 no junction control, 1 otherwise
2nd Road Class	0 TO 6	0 motoryway, 1 othewise
2nd Road Number	Numeric	exclude
Pedestrian Crossing-Human Control	0 TO 2	0 no pedestrain crossing, 1 otherwise
Pedestrian Crossing-Physical Facilities	0 TO 8	0 no pedestrain crossing, 1 otherwise
Light Conditions	1 TO 7	0 daylight 1 otherwise
Weather Conditions	1 to 9	0 good conditions, 1 otherwise
Road Surface Conditions	1 to 7	0 dry, 1 otherwise
Special Conditions at Site	0 to 7	0 no special conditions, 1 otherwise
Carriageway Hazards	0 to 7	0 no hazard, 1 otherwise
Urban or Rural Area	1 to 3	0 urban, 1 otherwise
Did Police Officer Attend Scene of Accident	1 TO 3	0 attend, 1 otherwise

TABLE II. VEHICLE FEATURES

Variable Name	Type	Preprocessing
Accident Index	Link field	EXECLUDE (unique accident, one or more vehicle)
Vehicle Reference	Link field	EXECLUDE (unique vehicle & one or more casualty)
Vehicle Type	1 TO 113	0 car 1 otherwise
Towing and Articulation	1 TO 5	0 no towing, 1 otherwise
Vehicle Manoeuvre	1 TO 18	0 reversing,1 otherwise
Vehicle Location-Restricted Lane	0 TO 10	0 on main c way, 1 otherwise
Junction Location	0 TO 8	0 no juntion, 1 otherwise
Skidding and Overturning	0 TO 5	0 no skidding, 1 otherwise
Hit Object in Carriageway	0 TO 12	0 no object, 1 otherwise
Vehicle Leaving Carriageway	0 TO 8	0 not leaving, 1 otherwise
Hit Object off Carriageway	0 TO 11	0 not object off c way, 1 otherwise
1st Point of Impact	0 TO 4	0 no impact, 1 otherwise
Was Vehicle Left Hand Drive	1 TO 2	0 right hand, 1 otherwise
Journey Purpose of Driver	1 TO 6	0 work, 1 otherwise
Sex of Driver	1 TO 3	0 male, 1 otherwise
Age Band of Driver	1 TO 11	
Engine Capacity	Numeric	
Vehicle Propulsion Code	1 TO 10	0 Petrol, 1 otherwise
Age of Vehicle (manufacture)	Numeric	
Driver IMD Decile	0 TO 10	0 deprived 1 othewise
Driver Home Area Type	1 TO 3	0 deprived 1 othewise

TABLE III. Casualty Features

Variable Name	Type	Preprocessing
Accident Index	Link field	EXECLUDE (unique accident, one or more casualty)
Vehicle Reference	Link field	EXECLUDE (unique in casualty table)
Casualty Reference	Link field	EXECLUDE (unique in vehicle and one or more in casualty)
Casualty Class	1 TO 3	0 driver, 1 otherwise
Sex of Casualty	1 TO 2	0 male, 1 otherwise
Age Band of Casualty	Numeric	SCALE
Casualty Severity	1 TO 3	TARGET VALBLE (0 fatal , 1 otherwise)
Pedestrian Location	1 TO 10	0 crossing,1 otherwise
Pedestrian Movement	1 TO 9	0 crossing,1 otherwise
Car Passenger	0 TO 2	0 passenger, 1 otherwise
Bus or Coach Passenger	0 TO 4	0 bus or coach passenger, 1 otherwise
Pedestrian Road Maintenance Worker (From 2011)	0 TO 2	0 road worker, 1 otherwise
Casualty Type	0 TO 113	0 pedestrain 1 otherwise
Casualty IMD Decile	0 TO 10	0 deprived 1 othewise
Casualty Home Area Type	1 TO 3	0 urban 1 otherwise

C. Data Analysis

Methods for traffic accident prediction can be broadly classified into three categories, namely statistical models, machine learning and analytics approaches, and simulation-based methods [17]. In this research we will concentrate on machine learning approaches.

Machine learning is a broad concept, which include supervised learning and unsupervised techniques. Supervised learning techniques include:artificial neural networks and its variations (Deep Learning, self-organised map), support vector machine (SVM), decision trees, Bayesian inference. Unsupervised learning include: association rules and clustering techniques.

Unsupervised learning involves searching for previously unknown patterns or groupings. Usually these techniques work without a prior target variable. Clustering and association rules fall under this group of techniques. Supervised learning, on the other hand, involves classification, prediction and estimation techniques that contain a target variable. Classification is a machine learning technique that assigns a class to an instance, i.e. automatically assigning traffic accident to one predefined class of severity. Prediction is similar to classification but involve assigning a continuous value to an instance.

Supervised learning methods usually use two sets: a training and test set. Training data is used for learning the model and requires a primary group of labelled traffic accident. Test set is used to measure the efficiency of the learned model and includes labelled traffic accident instances, which do not participate in learning classifiers.

This paper focuses on applying classification methods to classify accident severity. Five techniques will be applied and compared: (1) logistic regression models, (2) deep neural networks, (3) support vector machines, (4) decision trees, (5) extreme gradient boosting.

1) Logistic Regression: Regression models have become an integral component of any data analysis concerned with the relationship between a response variable and one or more explanatory variables. Logistic regression is a maximum-likelihood method that has been used in hundreds of studies of crash outcome.Traditionally, statistical regression models are developed in highway safety studies to associate crash frequency with the most significant variables. The logistic regression is a special case of the generalized linear model (GLM), which generalizes the ordinary linear regression by allowing the linear model to be related with a response variable that follows the exponential family via an appropriate link function. Logistic regression can be binomial or multinomial. The binomial logistic regression model has the following form:

$$p(y|x, w) = Ber(y|sigm(w^T x))$$

where w and x are extended vectors, i.e.,

$$w = (b, w1, w2, ..., w_D), x = (1, x1, x2, ..., x_D).$$

2) Artificial Neural Networks: Artificial Neural Networks (ANN) was build to imitate how the human brain works. It is formed by creating a network of small processing units called Neurons. Each neuron is very primitive, but the network can achieve complex tasks such as pattern recognition, image classification, and detection, natural language processing, etc. Mathematically ANN can be looked like a type of regression system that predicts and estimates new values from historical records. ANN is able to estimate any non-linear functions provided enough datasets were supplied for training the ANN. The architecture of ANN is built with three layers:

1) Input Layer: This layer receives the feature for the model.
2) Hidden Layer: This layer consists of one or more layers that identify the depth of the ANN. Each layer is connected through nodes with weighted edges. The performance of the model depends greatly on the hidden layers and their connectivity with input and output layers.
3) content...

3) Support Vector Machines: Support Vector Machines (SVMs) have been introduced as a new and novel machine learning technique according to the statistical learning theory. SVMs are used for classification and regression problems. Structural Risk Minimization (SRM) applied by SVM can be superior to Empirical Risk Minimization (ERM) since SRM minimizes the generalization error.

The primal form of SVM for classification is:

$$H : y = f(x) = sign(wx + b)$$

For regression the SVM is represented as:

$$H : y = f(x) = w^T x + b$$

4) Decision Trees: Decision trees are powerful data mining methods that can be used for classification and prediction.Decision trees represent rules, which are easy to interpret. There are a multiple methods used in creating decision trees for example: Iterative Dichotomiser 3 (ID3) and C4.5. Decision trees are supervised learning methods. The decision trees working mechanism is to divide the data data into training and testing sets randomly.

Decision trees have high variance because the model yields to different results. Namely, bagging and boosting. In Bagging techniques, many decision trees build in parallel, form the base learners of bagging technique.The sampled data is input to the learners for training.

In boosting techniques, the trees are build sequentially with fewer splits. Such small trees, which are not very deep, are highly interpretable. The validation techniques like k-fold helps in finding the optimal parameters which helps in finding the optimal depth of the tree. Also, it is very important to carefully stop the boosting criteria to avoid over-fitting.

5) eXtreme Gradient Boosting (XGBoost): XGBoost, a scalable machine learning system for tree boosting which is proofed very popular in machine learning competitions such as kaggle and kdnuggests Most winning teams either utilize or supplement their solution with XGBoost . This success can be mainly attributed to the scalabitliy feature that is inherit inside the algorithm. Scalabitliy is due to the optimized learning algorithm to work with sparse data,parrallisim and its utilitisation of mutlithreading [20]. XGBoost is a boosting algorithm which uses gradient descent optimization technique with a regulized learning objective function. It has the following features:

1) Regularization: XGBoost prevents overfitting by using L1 and L2 regularization.
2) Weighted quantile sketch: Finding the split points is core task of most decision tree algorithms. Their performance affected if the data is weighted. XGBoost handles weighted data through a distributed weighted quantile sketch algorithm.
3) Block structure for parallel learning: XGBoost utilizes multiple cores on the CPU using a block structure which part of its design. Data is sorted and stored in in-memory units or blocks which enables the reuse of data by iterations. This also useful for split finding and column sub-sampling tasks.
4) Handling sparse data: Data can become sparse for many reasons such as missing values or one-hot encoding. XGBoost split finding algorithm can handle different types of sparsity patterns in the data.
5) Cache awareness: In XGBoost, non-continuous memory access is required to get the gradient statistics by row index. Hence, XGBoost has been designed to make optimal use of hardware. This is done by allocating internal buffers in each thread, where the gradient statistics can be stored.
6) Out-of-core computing: This feature optimizes the available disk space and maximizes its usage when handling huge datasets that do not fit into memory.

D. Model Evaluation

Evaluation is a key stage in the data analytics that assesses the predictive capability of the model and identify the model which performs best [17]. Several techniques normally used to evaluate classification models such as the confusion matrix, receiver operator curve (ROC) and the area under the curve (AUC). A confusion matrix shows the correct classifications true positives (TP) and true negatives (TN) in addition to incorrect classification false positives (FP), and false negatives (FN) [21]. The accuracy is calculated from the confusion matrix which gives the precision (percentage of data correctly classified) and recall (percentage of data which are correctly labelled) values. The equations from 1-6 shows the performance matrices formulas.

$$TPR = \frac{TP}{TP + FN} \tag{1}$$

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F - measure = \frac{2 * Recall * Precision}{recall + precision} \tag{6}$$

V. Results and Analysis

Using python, jupyter notebook and Scikit learn, pandas and matplot data science libraries we have developed a workflow for processing the dataset and generate the corresponding accident severity prediction models. It is composed of a number of nodes, namely:

1) Dataset: contains the pre-processed data for the experiment
2) Explore Data: is an optional node to help in data exploration and viewing some statistics about the data before modelling.
3) Model: contains the algorithms that will be used for model generation.
4) Apply: where the model is applied to the predictors to generate the required results
5) Predictors: sample dataset for testing the prediction.
6) Prediction: the resulted table after applying the model on the predictors.

The dataset we have used is the UK traffic accidents that occurred between 2005-2019 obtained from UK department of transport. The data comes in three different files:accidents, casualties and vehicles in Tables I, II and III. Accidents_ index field joins the three tables in a one to many relationship, with one accident record corresponds to one or more casualties and vehicle records. Casualty table has a field called Vehicle Reference links a particular casualty record with vehicle and driver information record with the same accident_ index field. The original dataset contains about 2M accidents, 2M casualties and 5M Vehicles. The combined table resulted in records 3M records.

Data was explored using bar charts and histograms to look for trends and patterns in the data. Examples of such graphs are shown in Fig. 2, 3 and 4.

The following observations can be noted from the data:

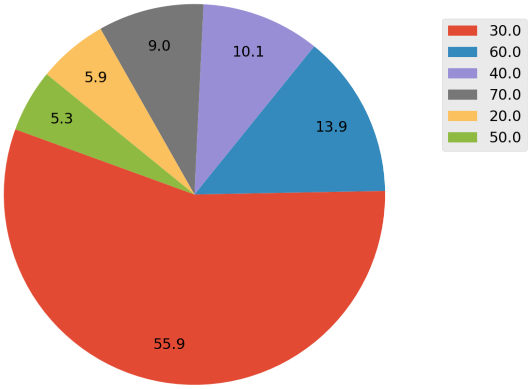


Fig. 2. Accidents per Speed Zone

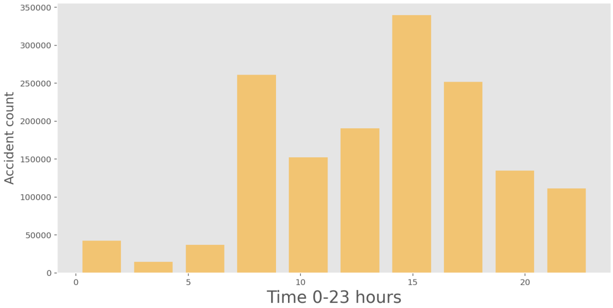


Fig. 3. Accident Time of Day

1) Accident severity (more than 90% records with slight severity).
2) Most crashes involve less than five cars with median 2 and mean 1.8.
3) Number of casualty range between 1 to 93 with mean 1.3 and median 1.
4) Accidents spread throughout the week with slight increase of accidents in Thursdays.
5) Accidents spread throughout the day with slight increase at school return time during working days.
6) First road class 3 roads has maximum number of accidents and single carriage ways roads and speed limit 30 MPH.
7) Uncontrolled junctions has more accidents than other types of accident.
8) Most accidents occur with fine weather conditions with Dry road surface.

The date then cleaned from incomplete records. All records with empty cells or value -1 were considered missing and removed. Then a histogram diagram was created for each column and non widespread columns were removed.

1) $Bus\_or\_Coach\_Passenger$
2) $Towing\_and\_Articulation$
3) $Vehicle\_LocationRestricted\_Lane$
4) $st\_Point\_of_Impact$
5) $Was\_Vehicle\_Left\_Hand\_Drive?$
6) $st\_Road\_Class, st\_Road\_Number$
7) nd_Road_Class

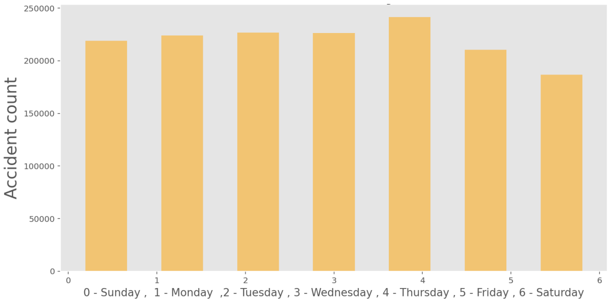


Fig. 4. Accidents Day of the Week

8) nd_Road_Number
9) $Pedestrian\_Crossing$
10) $Human\_Control$
11) $Special\_Conditions\_at_Site$

The resulting number of attributes used for model building is 48 features. Extra preprocessing was implemented on category type attributes by encoding with 0 and 1 values. Numeric and ordinal fields were scaled to remove bias due to large values. Out of 3M records of accidents only 29K were fatalities and 190K serious injuries. Serious and fatal records are grouped into one class and slight injury was the second class. This reduces the class imbalance together with under sampling method will enables the models achieve better results.

The selected models for this experiment are: logistic regression, decision trees, random forest, neural networks and XGBoost. Hyper parameter tuning was applied the methods. The dataset was partitioned two parts, 70% for training and 30% as test data. The metrics used for evaluating the algorithms were: balanced accuracy which is usually used with imbalanced data set. The balanced accuracy results is shown Table IV and ROC curves are shown in Fig. 5.

TABLE IV. BALANCED ACCURACY RESULTS

Method	balanced accuracy
Logistic Regression	66.26
Decision Trees	69.42
Support Vector Machines	53.22
Neural Networks	67.23
Random Forest	73.82
XGBoost	74.40

XGBoost and Random forest has shown better performance than logistic regression and, support vector machine and neural networks. This can be attributed to the nature of the modelling task and the data used. Most attributes have category values where decision trees based methods are reported to outperform regression based methods. Although with high number of dimensions decision trees based methods tend to affect its performance, in this data these methods continue to outperform linear and non linear classifiers. One downside is performance, however, as the data size increases time increased to obtain the results compared to logistic regression. In addition, further investigations needed to be undertaken to compare the performance with rule based methods which are report to perform well with categorical data.

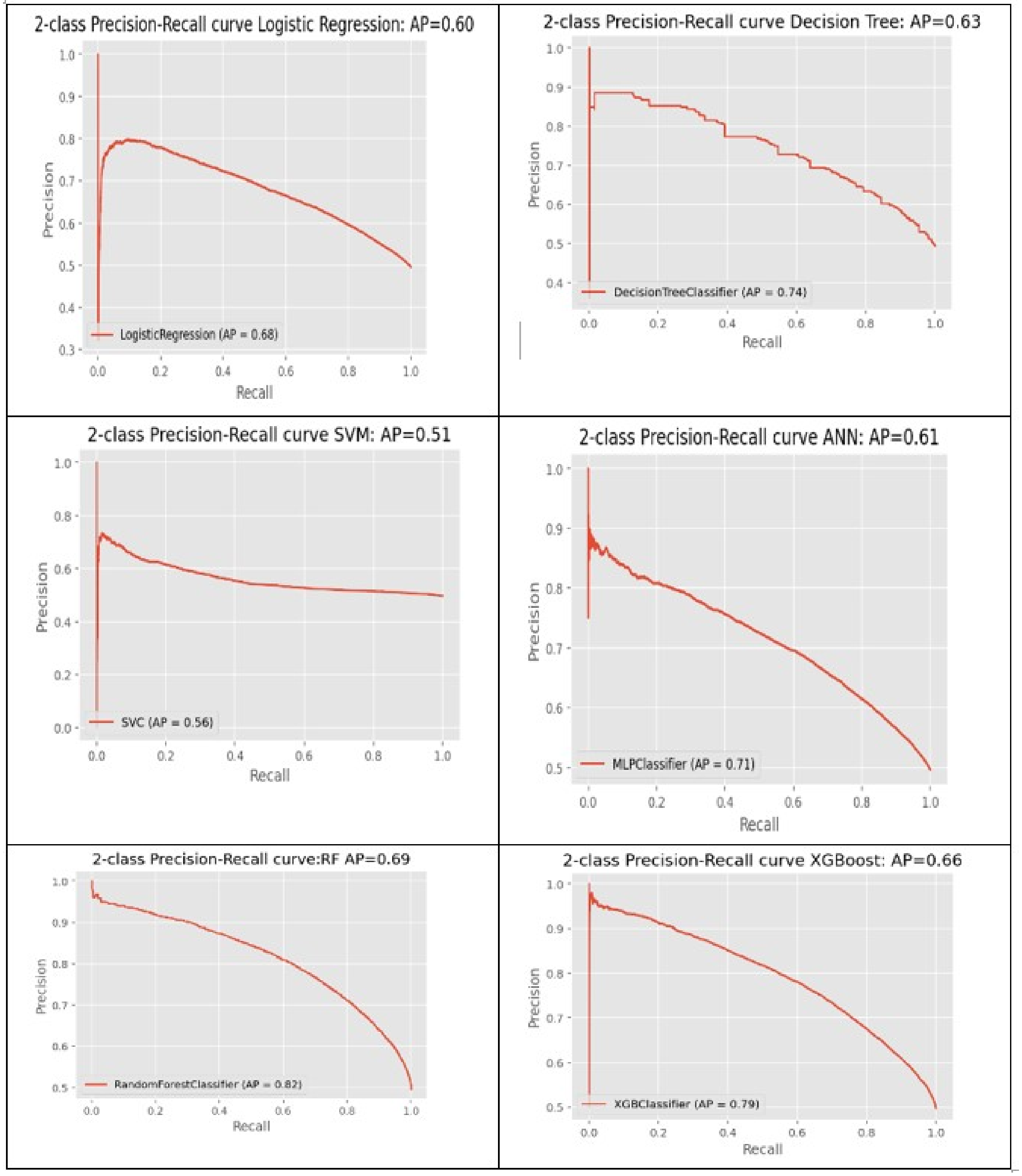


Fig. 5. ROC Curves for LR, SVM, ANN, DT, RF, XGBoost

The feature importance figure can be shown in Fig. 6. The figure shows the top 20 attributes that has effect on severity level. The top attribute is the casualty type which specify whether the casualty was a pedestrian or a passenger. This is followed by vehicle and area attributes. Although 66% of accidents were in 30 miles speed limit , it appear from the feature importance table that speed limit has less effect on the injury type. This insight can help raod traffic authorities to prioritise measures to reduce injury levels.

VI. Conclusion

This paper presented a data analytic framework in which UK traffic accidents data was analysed to established a model for predicting injury severity. The paper used publicly available data from 2005 to 2019 to build prediction models for injury severity level. The paper has combined all attributes from three data sources to analyse 63 attributes and it relation with accident severity. The paper highlighted issues related to data quality and imbalanced data and applied techniques to tackle these issues. The paper compared performance between

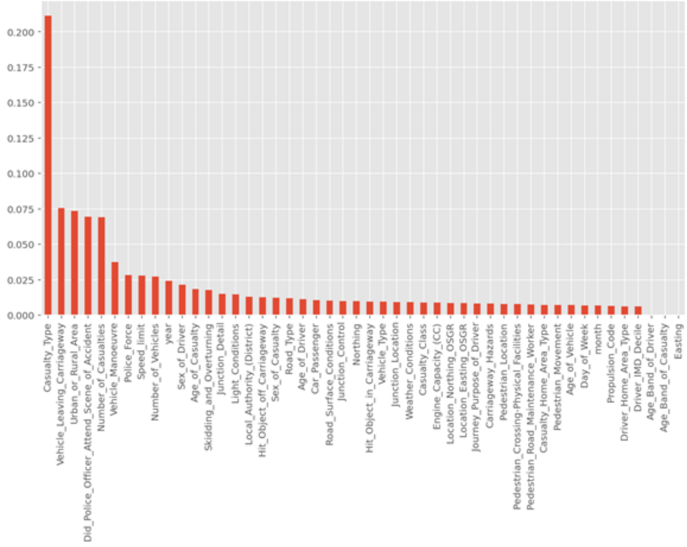


Fig. 6. Feature Importance

different machine learning techniques, XGBoost algorithm was shown to outperform other techniques with higher accuracy rate even with imbalanced data. Further work is suggested to use parallel processing libraries and compare the performance of rule based techniques and decision based techniques.

Acknowledgment

This work is supported by grant number 17-COM-1-01-0007,Deanship of Scientific Research (DSR) of Umm al Qura University, Kingdom of Saudi Arabia. The authors would like to express their gratitude for the support and generous contribution towards pursuing research in this area

References

[1] World Health Organization (WHO), *A Road Safety Technical Package*, 2017. [Online]. Available: http://iris.paho.org/xmlui/bitstream/handle/123456789/34980/9789275320013-por.pdf?sequence=1{\&}isAllowed=y

[2] A. Mehdizadeh, M. Cai, Q. Hu, M. A. A. Yazdi, N. Mohabbati-Kalejahi, A. Vinel, S. E. Rigdon, K. C. Davis, and F. M. Megahed, "A review of data analytic applications in road traffic safety. Part 1: Descriptive and predictive modeling," *Sensors (Switzerland)*, vol. 20, no. 4, pp. 1–24, 2020.

[3] Q. Hu, M. Cai, N. Mohabbati-Kalejahi, A. Mehdizadeh, M. A. A. Yazdi, A. Vinel, S. E. Rigdon, K. C. Davis, and F. M. Megahed, "A review of data analytic applications in road traffic safety. Part 2: Prescriptive modeling," *Sensors (Switzerland)*, vol. 20, no. 4, pp. 1–19, 2020.

[4] A. Ziakopoulos and G. Yannis, "A review of spatial approaches in road safety," *Accid. Anal. Prev.*, vol. 135, no. July, p. 105323, 2020. [Online]. Available: https://doi.org/10.1016/j.aap.2019.105323

[5] S. Moosavi, M. H. Samavatian, A. Nandi, S. Parthasarathy, and R. Ramnath, "Short and long-term pattern discovery over large-scale geo-spatiotemporal data," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 2905–2913, 2019.

[6] N. Zagorodnikh, A. Novikov, and A. Yastrebkov, "Algorithm and software for identifying accident-prone road sections," *Transp. Res. Procedia*, vol. 36, pp. 817–825, 2018. [Online]. Available: https://doi.org/10.1016/j.trpro.2018.12.074

[7] D. W. Kononen, C. A. Flannagan, and S. C. Wang, "Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes," *Accid. Anal. Prev.*, vol. 43, no. 1, pp. 112–122, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.aap.2010.07.018

[8] D. Delen, R. Sharda, and M. Bessonov, "Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks," *Accid. Anal. Prev.*, vol. 38, no. 3, pp. 434–444, 2006.

[9] A. Naseer, M. K. Nour, and B. Y. Alkazemi, "Towards deep learning based traffic accident analysis," in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, 2020, pp. 0817–0820.

[10] B. Sharma, V. K. Katiyar, and K. Kumar, "Traf fi c Accident Prediction Model Using Support Vector Machines with Gaussian Kernel Á Accident characteristics Á Data mining Á," *Adv. Intell. Syst. Comput.*, vol. 437, pp. 1–10, 2016.

[11] H. Meng, X. Wang, and X. Wang, "Expressway crash prediction based on traffic big data," *ACM Int. Conf. Proceeding Ser.*, pp. 11–16, 2018.

[12] M. Schlögl, R. Stütz, G. Laaha, and M. Melcher, "A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset," *Accid. Anal. Prev.*, vol. 127, no. January, pp. 134–149, 2019. [Online]. Available: https://doi.org/10.1016/j.aap.2019.02.008

[13] J. Ma, Y. Ding, J. C. Cheng, Y. Tan, V. J. Gan, and J. Zhang, "Analyzing the Leading Causes of Traffic Fatalities Using XGBoost and Grid-Based Analysis: A City Management Perspective," *IEEE Access*, vol. 7, pp. 148 059–148 072, 2019.

[14] L. G. Cuenca, E. Puertas, N. Aliane, and J. F. Andres, "Traffic Accidents Classification and Injury Severity Prediction," in *2018 3rd IEEE Int. Conf. Intell. Transp. Eng. ICITE 2018*, 2018, pp. 52–57.

[15] Department for Transport, "Road Traffic Statisticsn guidance," pp. 1–13, 2014. [Online]. Available: http://data.dft.gov.uk/gb-traffic-matrix/all-traffic-data-metadata.pdf

[16] B. Burdett, "Improving Accuracy of KABCO Injury Severity Assessment by Law Enforcement," *Univ. Wisconsin-Madison*, 2014.

[17] J. Han, M. Kamber, and J. Pei. (2012) Data mining concepts and techniques, third edition. Waltham, Mass. [Online]. Available: {http://www.amazon.de/Data-Mining-Concepts-Techniques-Management/dp/0123814790/ref=tmm_hrd_title_0?ie=UTF8\&qid=1366039033\&sr=1-1}

[18] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*, 2018.

[19] L. Gautheron, A. Habrard, E. Morvant, and M. Sebban, "Metric learning from imbalanced data," *Proc. - Int. Conf. Tools with Artif. Intell. ICTAI*, vol. 2019-Novem, no. 9, pp. 923–930, 2019.

[20] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-August-2016, pp. 785–794, 2016.

[21] P. Hájek, M. Holeňa, and J. Rauch, "The GUHA method and its meaning for data mining," *J. Comput. Syst. Sci.*, vol. 76, no. 1, pp. 34–48, 2010.

Analysis of National Cybersecurity Strategies

Alexandra Santisteban Santisteban[1]
Lilian Ocares Cunyarachi[2], Laberiano Andrade-Arenas[3]
Facultad de Ciencias e Ingeniería
Universidad de Ciencias y Humanidades

Abstract—**Nowadays the use of information and communication technology has been incorporated in a general way in the daily life of a nation allowing the optimization in its processes. However, with it comes serious risks and threats that can affect cyber security because of the vulnerability they show. In addition, there are several factors that contribute to the proliferation of criminal actions in cyber security, the profitability offered by its exploitation in economic, political or other terms, the ease and low cost of the tools used to carry out attacks and the ease of hiding the attacker, make it possible for these activities to be carried out anonymously, from anywhere in the world and with impunity.The main objective of the research is to analyze and design National Cybersecurity Strategies to counter attacks. The methodology of this research was conducted in an exploratory and descriptive manner. As a result of the research work, a design of National Cybersecurity Strategies was obtained after an in-depth analysis of the appropriate strategies and thus minimizing the different attacks that can be carried out.**

Keywords*—*Cybersecurity; national strategies; risks; threats; vulnerability

I. Introduction

Technology continues to play a profound role in the global risk panorama, it is an issue that affects all of society which should not be treated as something insubstantial i.e. because of the large number of victims as a result of cyber-attacks, who are not aware of the risks that are exposed when they do not take certain knowledge into account [1].

Concerns about data fraud and cyber attacks is a very latent issue worldwide [2]. Today according to the General Packet Radio Service highlights a number of technological vulnerabilities about two-thirds of respondents expect risks associated with false news and identity theft, while three-fifths said the same about loss of privacy for businesses and governments [3].

Consequently, a security strategy can be seen as a key element in a nation's cybersecurity, which can help improve the resilience of national information infrastructures and services [4]. A strategy is established at a high level in the hierarchical structure of a nation, which sets out a series of national objectives and priorities to be achieved within a given time frame. As such, it provides a strategic framework for a nation's cybersecurity efforts [5]. While the tools, attacks and risks may be universal, the strategies are changing according to the policies adopted by different countries or groups of countries, as for example the European Union bases its strategies on the privacy of data or information, gives a context of principles, ethics, to safeguard the universal right to privacy [6]. In Latin America, most States have the capacity to respond to cyber-attacks, but the truth is that only six have designed a Cybersecurity Strategy.

The last one to present its Strategy was Mexico, on November 13, 2017, joining the small group of Latin American countries that, according to the OAS, have this type of policy; the rest are Colombia, Panama, Paraguay, Chile and Costa Rica [7]. Although the Republic of Peru does not have a National Policy on the subject, this year the law on cybersecurity was approved, with the aim of providing a legislative framework on cybersecurity in the country and the law on cyberdefense, which seeks to provide a regulatory framework for cyberdefense considering its capabilities as the development and implementation of military operations in cyberspace. During the law of cyber defense, it is mentioned that the Joint Command of the Armed Forces is responsible for monitoring and implementing cyber defense plans [8].

It is important for every country to have national policies and strategies and a plan of response to possible risks that may occur in the nations and thus be in the forefront of possible attacks. With national policies and strategies and a plan to respond to possible risks, can attacks be mitigated?

The objective of the research is to analyze and design national cybersecurity strategies in order to have a prevention alternative to possible attacks.

This paper is structured as follows: Section II will describe the methodology in detail. Section III will show the results and discussions obtained and finally, Section IV will present the conclusions according to the objective set.

II. Methodology

The article is an exploratory-descriptive research focused on the national cybersecurity strategy

A. Analysis of National Cybersecurity Strategies

1) Principles of a cybernetic strategy : A cybersecurity strategy must have a clear set of principles that provide a framework for decision making in the identification, management and mitigation of security risks. A cybersecurity strategy must have basic principles where there is a balance between civil rights, the right to privacy, costs and other priorities.

Table I mentions the national strategy by sector, operational, technical which is explained in detail, the development of a cyber security strategy focuses on the identification, analysis and evaluation of risks to be managed. Risks in cyberspace are typically thought of as risks to information

TABLE I. CYBER SECURITY STRATEGIES

National Strategy (Defense)	Best practices Standars,technology, process, people
Strategy By Sector	Components: O.S + Internet + Servers involved
Operative	Application Method
Technical	functions: E-Commerce, Crime, CII, Others

systems that if exploited, could negatively impact the economic well-being of the city or the public security of its citizens to a significant degree [7].

2) National Action Plan: Any state must have a comprehensive digital security plan that is part of a larger national security plan. Governments must be clear that the purpose of cybersecurity is to help preserve the organizational, human, financial, technological, and information resources necessary to achieve their goals [8]. The purposes of the security of a country, has to focus basically and mainly on the following points:

- Reduce vulnerabilities and threats.
- Limit the damage or dysfunction that could be induced by a security breach.
- Every nation has a plan of action that goes from the general to the particular.

3) Cybersecurity Strategies and Structures: For cybersecurity strategies to be sound and effective, there must be political will on the one hand, and on the other hand, organizational structures must be able to adapt and respond to the specific needs of a nation. Political will is important so that the various plans that may be put forward for cybersecurity can be addressed by the various agencies of a government, which must also include competent people who are proactive and have the capacity to respond reactively in the accepted time frame [9].

Fig. 1 shows us the structure of cybersecurity, related to each of the mentioned cycles, where every structure must be designed in such a way that efficiency prevails.

4) National cyber security threats using threat modeling: Assess national cyber security threats using threat modeling. Threat modeling can help identify the assets that the city is trying to protect, as well as what it wants to protect them from. A threat model takes an inventory of key municipal assets and their threats, determines the likelihood that those assets need protection, looks at the city's ability to defend against threats, and determines the consequences of inaction. This approach allows city leaders to identify and mitigate potential security problems early, while the problems are relatively easy and inexpensive to solve. Categorizing threats online as shown in the table below can facilitate the assessment of threats and then develop specific preventive and reactive strategies [10].

Table II mentions the assessment of threats in a specific way in which they are divided in a passive and active way in order to be able to categorize threats online and to facilitate the assessment of threats in order to then develop specific preventive and reactive strategies [11].

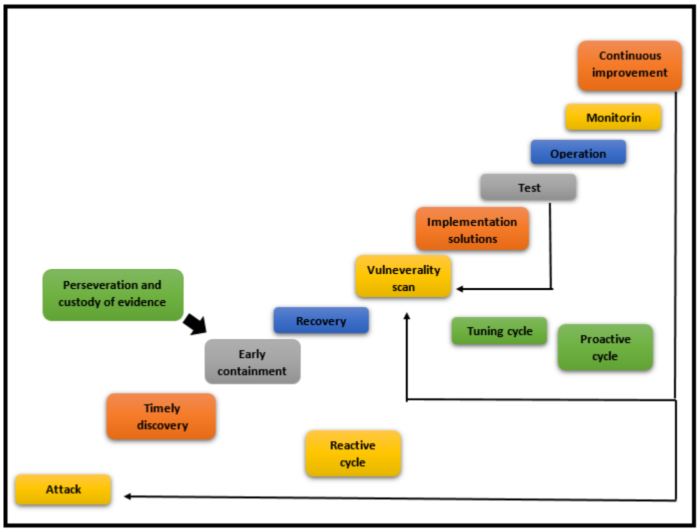


Fig. 1. Structure of Cybersecurity

TABLE II. THREAT ASSESSMENT

	Threat	Examples
Passives	Involuntary Actions	exposure to malware via email or websites
	Insufficient resources	Unprotected systems Unclear mitigation strategies Indefinite responsiveness Unclear membership
Active	Cyber crime	Fraud. Denial of service attacks. Theft of intellectual or financial property. Abuse or damage to TIC systems
	Natural hazards	Typhoons and hurricanes Earthquakes and tsunamis Floods Accidental cutting of submarine internet cables

5) Basic implementation of capabilities: Following the plan it becomes necessary to create capacities that serve as support and should be based on:

- Understanding the role of cybersecurity actors including their motivation, correlation, tools, mode of action, among others.
- The relevant generic safety functions of any safety action [12].All this will facilitate the identification of organizational structures to be effective and determine what kind of tools, knowledge and procedures should be effective to help solve cyber security problems and there are two main processes to be carried out.

The cybersecurity actors will be classified in the following points:

- The protector (private and/or public institutions).
- The one to be protected (the individual (citizen), the organization and the State) [13].
- Whether the criminal is professional or not.

TABLE III. PROACTIVE ACTIONS

They are based on:	A good understanding of TIC-related risks
	Technical, legal measures and complementary organization
	Effective security approachesand TIC quality management

Table III mentions technical measures and security approaches through ICT quality management. This method analyzes threats in addition to developing secure data. We make sure to consider important factors such as valuable data collection, memory storage, retrieval and a well-organized high-level network source to establish an intelligent city [14].

- Reducing the number of vulnerabilities, the number of potential targets and their interconnection would help to create an environment that is difficult to be vulnered.
- The levels of risk perception must be increased in order to observe in detail and avoid future problems and also to decrease the expected benefits [15].

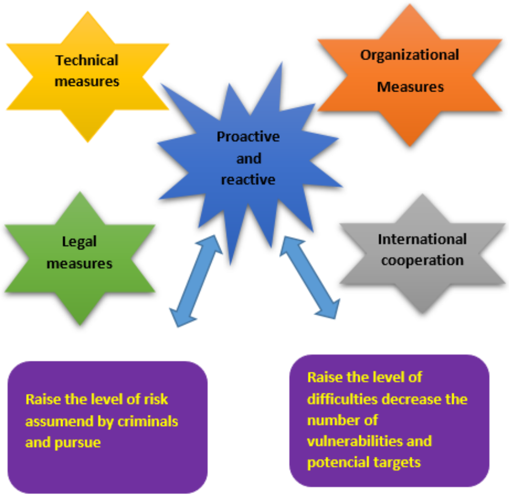

Technical measures
Organizational Measures
Proactive and reactive
Legal measures
International cooperation
Raise the level of risk assumend by criminals and pursue
Raise the level of difficulties decrease the number of vulnerabilities and potencial targets

Fig. 2. Components of the Security Strategy

Fig. 2 shows us the components of the security strategy where investigating, identifying and responding to online threats must be a primary component of the cyber security strategy.

To achieve these strategic objectives of protection must be implemented information and communication security solutions such as raising the level of effort required to carry out an attack, makes the potential specific resources can be less vulnerable if the robust security solutions are well designed, implemented and managed. With an implementation of a network security architecture, through the use of access controls, integrity or authentication or through surveillance mechanisms, with these measures attacks become more difficult to carry out, and this in turn leads to a reduction in incidents. In the face of this, legislative and regulatory measures must strictly seek to support or contribute to increasing the level of perceived risk, and reduce the favorable context for perpetrating an illegal action [15].

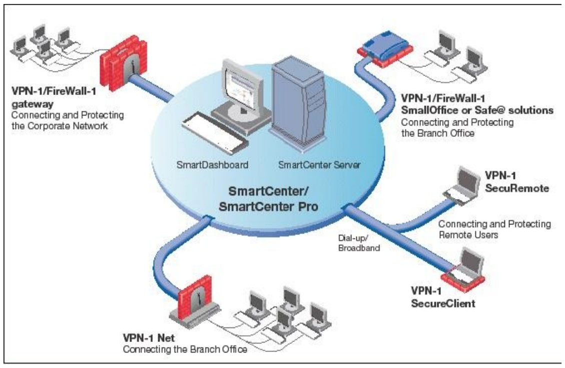

VPN-1/FireWall-1 gateway
Connecting and Protecting the Corporate Network
VPN-1/FireWall-1 SmallOffice or Safe@ solutions
Connecting and Protecting the Branch Office
SmartDashboard
SmartCenter Server
SmartCenter/ SmartCenter Pro
VPN-1 SecuRemote
Connecting and Protecting Remote Users
Dial-up/ Broadband
VPN-1 SecureClient
VPN-1 Net
Connecting the Branch Office

Fig. 3. Network Security Architecture

In Fig. 3, the security architecture of the network is mentioned, through the use of access controls, integrity or authentication or through surveillance mechanisms. This architecture constitutes a general, simple and flexible working model for the tasks of Planning, Implementation and Maintenance of security, which integrates a group of components that consider the most important aspects inherent to network security.

6) Comparison of Cybersecurity Strategies: For this work we selected a group of countries that have developed and published their National Cybersecurity Strategies in order to analyze and make a comparison between them, so we will take into account the main aspects on which most countries are focused that have implemented the European cyber security network and should be taken into account in the development of strategies or policies for cybersecurity, to address the risks of cyberspace [16].

6.1 Information Protection in Cybersecurity

Critical infrastructure: Refers to the set of computers, computer systems, telecommunications networks, data and information, whose destruction or interference can weaken or impact the security of a nation [17].

Economy: refers to the presence of the economy in cyberspace, it is a way to organize the exchange of goods and services business-to-business, business-to-customer, regardless of geographical location, or time differences.

National Security: notion of relative stability, calm and security, beneficial for the development of a country, as well as the implementation of resources and strategies to achieve it.

Social Welfare: set of factors that participate in the quality of life of people and make their existence has all those elements that give rise to the tranquility and human satisfaction.

6.2 Strategy/Policy Focus

Governments and international organizations around the world have begun to develop specific cybersecurity strategies to address emerging threats in and from cyberspace. A new generation of government policies on cybersecurity has taken shape in several countries, including our case study from Austria, where a national cybersecurity strategy was developed. These new policies are characterized by similar strategic objectives and areas of focus, such as increasing reliance on public-private partnerships and international cooperation, along with major reforms in government structures [18].

Awareness: It is done with the aim of making society aware of individual risks (privacy and intimacy) and collective risks (national security, economic, social and cultural prosperity) that derive from an inadequate use of cyberspace.

Knowledge: Advanced knowledge of technology and the state of cyberspace must be maintained, and technological watch must be established in the area of cybersecurity to ensure that knowledge is obtained and cooperation projects promoted to achieve integration and maximum use of international opportunities, resources and advances.

Education: incorporating courses related to cybersecurity in education plans should be implemented from primary to higher education. The aim of initiating education at an early age is, on the one hand, to homogenize knowledge in the use of new technologies, as well as their responsible use and, on the other hand, to identify future cybertalents.

Military cyber capabilities: The ability of a country's armed forces to prevent and counter any threat or incident of a cyber nature that affects national sovereignty.

6.3 Public Sector Participation in Strategy/Policy

Leadership: the scope and complexity of the challenges of cyberspace require, in addition to national leadership, the appropriate coordination of the capacities, resources, and competencies involved. Both of these requirements must be assumed by the government, which will direct and oversee the National Cybersecurity Strategy/Policy.

Legal framework: have a strong legislative framework in the area of cybersecurity, which addresses the different types of crimes both nationally and internationally.

Leadership: the scope and complexity of the challenges of cyberspace require, in addition to national leadership, the appropriate coordination of the capacities, resources, and competencies involved. Both of these requirements must be assumed by the government, which will direct and oversee the National Cybersecurity Strategy/Policy.

6.4 Private Sector Participation

Participation in strategy/policy: actors in key sectors such as energy, transport, financial institutions, stock exchanges, internet service providers, among others, must assess the risks that affect them and through proper management of these risks, ensure that information systems and networks are reliable and resilient.

6.5 International Cooperation

Cooperation in your group: a geopolitical bloc is called a group in this case, which is the distribution formed by countries that share a certain extension of land, economic, political and cultural panorama.

Cooperation with other countries: technological globalization, its opportunities and its risks make it necessary to align the initiatives of all the countries that pursue a safe and reliable cyberspace.

B. Designing National Cybersecurity Strategies

This section describes the various phases in the development of a strategy, which are as follows:

1) Phase I - Initiation: The initiation phase of the national cybersecurity strategy lays the foundation for its efficient development. It is expected that this phase will focus on the processes, timelines, and identification of key stakeholders to be involved in the development of the strategy. This phase culminates in the development of a strategy preparation plan. When the country's administrative procedure so provides [19].

2) Phase II - Inventory and Analysis: The objective of this phase is to collect data to assess the national cybersecurity landscape and the current and future status of cybersecurity risks in order to obtain information for the purpose of drafting and developing the national cybersecurity strategy.

3) Phase III - Production: In this phase, the strategy text is developed with the participation of key stakeholders from the public sector, the private sector and civil society through a series of public consultations and working groups. This broader group of stakeholders, coordinated by the project authority, will be responsible for defining the overall vision and scope of the strategy [20].

4) Phase IV - Execution: The implementation phase is the most important of the entire life cycle of the national cybersecurity strategy. A structured approach to implementation, with adequate human and financial resources, is critical to the success of the strategy and should be considered part of its development. The implementation phase is usually based on an Action Plan, which guides the various activities planned.

5) Phase V - Monitoring and Evaluation: At this last stage, the competent authority should devise a formal process for monitoring and evaluating the strategy. In the monitoring phase, the government should ensure that the strategy is implemented in accordance with its Action Plan. In the assessment phase, the government and its competent authority should determine whether the strategy remains relevant in light of evolving risks, whether it continues to meet the government's objectives, and what adjustments are needed.

Fig. 4 shows the steps a nation must follow to develop a national strategy and the possible mechanisms for its implementation according to its specific needs and requirements, integrating general principles and good practices.

III. Results and Discussions

A. Analysis of the Ranking of the Most Attacked Countries

As a result of the 20 most attacked countries, we used Kaspersky's web page which allows us to consult the most attacked countries around the world in real time. The results achieved were from 15/10/2020 at 9:00 PM. Since this page updates the attacks per second, the data obtained is based on this, and the most attacked country is Russia. It was observed that this country does not easily change positions in the ranking, as it is the first country with cyber-attacks, and Brazil is in second place if it varies from position to position with

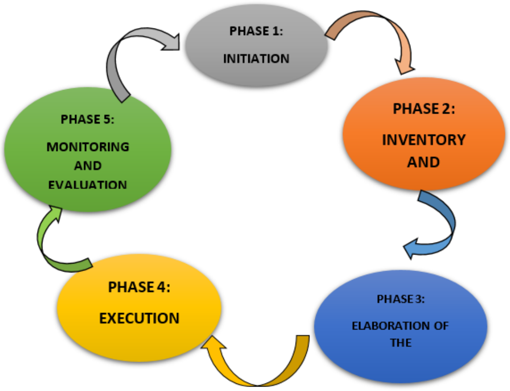


Fig. 4. Life Cycle of the National Cybersecurity Strategy

Germany, which is in third place, and so on until it reaches the ranking of 20. The generation of authentic variants of a specific malware results in a valid database of malware variants, which is searched by anti-malware scanners to identify the variants before they are released by the malware developers. This research employs a code avoidance strategy i.e. insertion and removal i.e. if available of a specific assembly code instruction that goes directly into the source code of the virus. Starting with a database of over 60 popular anti-virus scanners, this variant-based approach to malware generation successfully evolves from the timid variants that evade over 97% of anti-virus scanners. The results of this research demonstrate the potential for malware generation and also open up avenues for further analysis [21].

TABLE IV. Ranking of the most attacked countries

Ranking	country
1	Russia
2	Brazil
3	Germany
4	Vietnam
5	China
6	United States
7	France
8	Mexico
9	Indonesia
10	India
11	Spain
12	Japan
13	Malaysia
14	Canada
15	Italy
16	Thailand
17	Philippines
18	Colombia
19	Ecuador
20	Peru

Table IV shows the ranking of the most attacked countries. Among the five most attacked countries are Russia, Brazil, Germany, Vietnam and China, which are the most infected by cyber security attacks. The study shows how a unique position within the ecosystem can lead a company to dominate the market. As a result, actions aimed at creating preferential conditions for company services can be interpreted as restricting competition by promoting a discriminatory environment and preventing software developers from entering the market through cyber security [22]. Table V shows in detail the meaning of each acronym shown below.

TABLE V. KASPERSKY

	Acronym
1	OAS: On-Access Scan shows the flow of malware detection in the process of scanning with On- Access.
2	ODS: On Demand Scanner shows the flow of malware detection while scanning under order, when the user manually selects the option ”Search for viruses” in the context menu.
3	MAV: Mail Anti-Virus is given to show the flow of malware detection during thethat new objects appear so to speak related in an email application. What Mav does is that it acts at the moment of arrival of the messages and calls Oas when saved to those added to a disk.
4	WAV: Web Anti-Virus shows the flow of malware detection during scanning Web Anti-Virus where an HTML page from a website is opened or a file is downloaded.
5	IDS: Intrusion Detection Scan shows the flow detection of network attacks.
6	˜VUL: Vulnerability Scan shows the flow of the detection of vulnerabilities.
7	KAS: Kaspersky Anti-Spam shows the suspicious and unwanted trade found by Kaspersky’s filtration technologies.
8	BAD: Botnet Activity Detection shows statistics on people’s IP addresses who are victims of cyber attacks.These statistics were acquired with the help from the DDoS intelligence system.

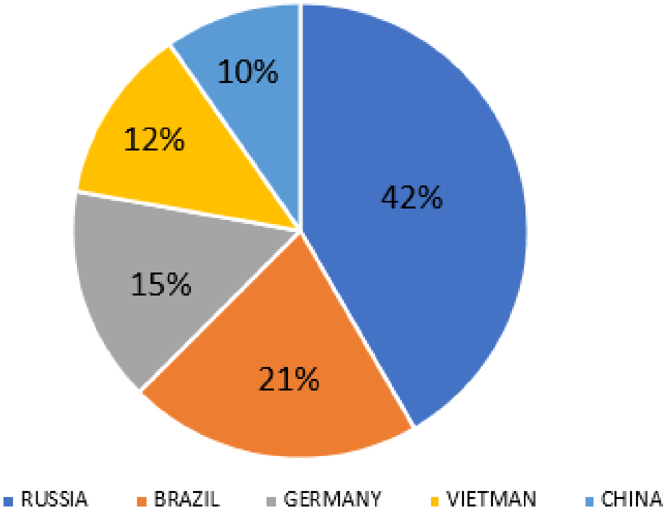


Fig. 5. Most Attacked Country

Fig. 5 shows the percentage of the top 5 most attacked countries as the first country is Russia with 42%, followed by Brazil with 21% as the third most attacked country is Germany with 15%, the fourth country is Vietnam with 12% and finally China with 10%.

In the Fig. 6 Russia is shown as the first country more attacked that has as data in On-access scan (142105), On demand scanner (31548), Mail anti virus (1424), Web anti virus (25297), Intrusion detection scan (512508), Vulnerability scan (687), Kaspersky anti spam (162609), Botnet activity detection (0).

Fig. 7 shows Brazil as the second most attacked country

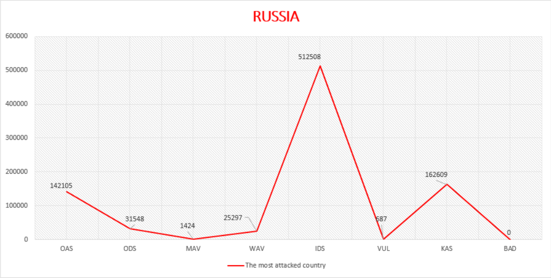


Fig. 6. First Most Attacked Country

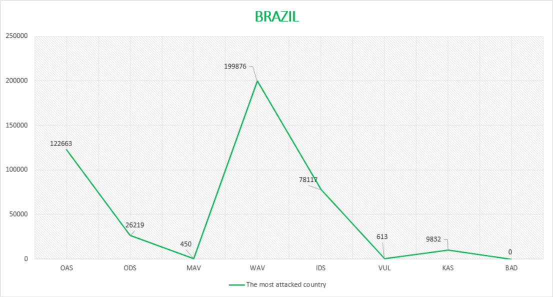


Fig. 7. Second Most Attacked Country

that has as data On-access scan (122663), On demand scanner (26219), Mail anti virus (450), Web anti virus (199876), Intrusion detection scan (78117), Vulnerability scan (613), Kaspersky anti spam (9832), Botnet activity detection (0).

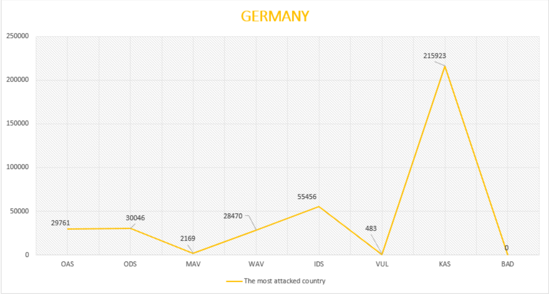


Fig. 8. Third Most Attacked Country

Fig. 8 shows Germany as the third most attacked country with On-access scan (29761), On demand scanner (30046), Mail anti virus (2169), Web anti virus (28470), Intrusion detection scan (55456), Vulnerability scan (483), Kaspersky anti spam (215923), Botnet activity detection (0).

Fig. 9 shows Vietnam as the fourth most attacked country with data in On-access scan (147117), On demand scanner (74015), Mail anti virus (6075), Web anti virus (21992), Intrusion detection scan (78992), Vulnerability scan (2263), Kaspersky anti spam (13371), Botnet activity detection (0).

Fig.10 shows China as the fifth most attacked country with data in On-access scan (60302), On demand scanner (18658), Mail anti virus (477), Web anti virus (3813), Intrusion detection scan (150522), Vulnerability scan (857), Kaspersky anti spam (55134), Botnet activity detection (0).

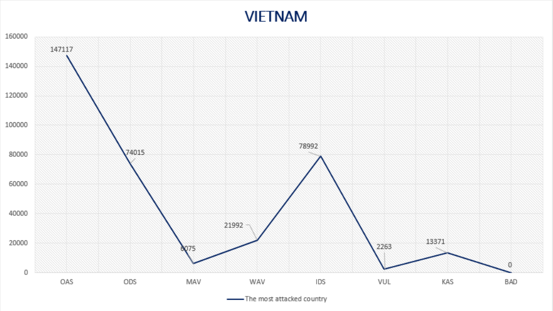


Fig. 9. Fourth Most Attacked Country

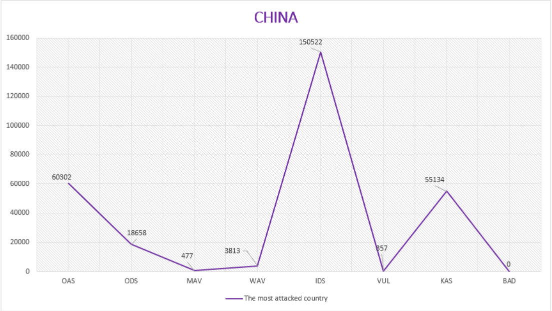


Fig. 10. Fifth Most Attacked Country

B. Comparison of Cyber Security Strategies

This includes countries that have a high cybersecurity rating. Cross-referencing these strategies will provide the necessary information on how the developing nations listed progressed at such a rapid pace, in the area of cybersecurity, leaving behind even many developed countries, the countries best positioned in each of the criteria set out above, most of which are European:

- Lowest percentage of infected mobile devices: Finland, 0.87% of users.
- Lowest number of financial malware attacks: Denmark, Ireland and Sweden, 0.1% of users.
- Lowest percentage of infected computers: Denmark, 3.15% of users.
- Lowest percentage of cyber attacks by country of origin: Turkmenistan, 0%.
- Country best prepared for cyber attacks: United Kingdom, score of 0.931 or 93.1%.
- Most up-to-date legislation to date: France, China, Russia and Germany have all seven categories covered [23].

Table VI shows the ranking of developing countries with high cyber security, which have extremely high cyber-crime rates, so analysis of their strategies will provide considerable indications for protecting cyberspace against various threats and attacks. Denmark is the safest country in the world in terms of cyber security, surpassing powers such as Japan, which fell four places from last year's ranking.

TABLE VI. Developing countries with high cyber security

Security ranking cybernetics	country
1	Denmark
2	Japan
3	france
4	Russia
5	Germany

Cyber threats are devastating. Billions are spent around the world to prevent relentless security attacks, but unless business and security are integrated and aligned, these threats will continue to exist and disrupt the operations of organizations [24]. The development of a comprehensive strategy can pose many challenges, as cooperation and agreement among stakeholders and a common course of action are needed, and this task will not be easy. It should be noted that the process of developing the strategy is likely to be as important as the final outcome document [25].

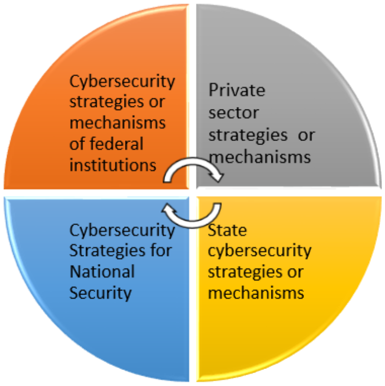

Cybersecurity strategies or mechanisms of federal institutions
Private sector strategies or mechanisms
Cybersecurity Strategies for National Security
State cybersecurity strategies or mechanisms

Fig. 11. Strategies in Different Environments

In the Fig. 11, it is shown each part in which it includes the strategy used in different environments in which they have to follow by different processes to reach the final result of the strategy.

TABLE VII. Cybersecurity Strategies

	Cybersecurity Strategies
1	Australia
2	Austria
3	Canada
4	Czech Republic
5	Estonia
6	Finland
7	France
8	Germany
9	India
10	Iran
11	Israel
12	Japan
13	Malaysia
14	Netherlands
15	New Zealand
16	Saudi Arabia
17	Spain
18	Turkey
19	United Kingdom
20	EE.UU

In Table VII, we identified the cybersecurity strategies we selected 20 countries in which Australia is positioned as the first country to carry out cybersecurity strategies in table 5 we identified the developing countries with high cybersecurity based on that the strategic points are made in different countries such as Austria, Canada, Czech Republic, Estonia, Finland, France, Germany, India, Iran, Israel, Japan, Malaysia, New Zealand, Saudi Arabia, Spain, Turkey, United Kingdom, USA. In the United States, these countries use strategies to combat cyber-crime at a global level. The national strategy for cybersecurity has been enshrined in the creation of the National Cybersecurity Forum to enhance and create public-private synergies, its implementation and the harmonization of its operation with existing parties, will be done through the adoption of the necessary regulatory provisions. Cybersecurity policies present spaces for political articulation and intervention where the very contours of an emerging digital society and the socio-technical relations of power and control that are considered necessary to govern its emergence are assembled. Globalizing form and rationality of security that codifies and enables new forms of control and intervention, but also new responsibilities at the interface between the State, society and individuals [26].

C. Proposal for a Strategic Cybersecurity Design

Fig. 12 shows the flowchart designed on the basis of the cybersecurity strategy, as it is fundamental to understanding whether the objectives of the strategy are being met or whether different actions should be taken. In this process, it is also necessary to periodically re-examine the overall risk context to understand whether external changes have occurred that may affect the strategy's outcomes. In the initiation phase of the national cybersecurity strategy in focuses on the processes and identification of one of the main sections of developing a strategy preparation plan, the strategy development plan should identify the main steps and activities, the most important parts, the time frame, and the required resources.

It should be determined how and when the parties should participate in the drafting process by giving their input and opinions. For the national cybersecurity strategy to be effective, it must demonstrate the country's position on cybersecurity. Indeed, an analysis of the country's existing cybersecurity strengths and weaknesses should be conducted, and key materials and documents should be consulted in cooperation with relevant authorities in the private sector government and civil society. Based on the information gathered in the previous phase, the project authority should assess the risks to which the country is exposed due to its digital dependency. This can be done by identifying the national public and private digital assets in addition to their interdependencies, weaknesses and threats, as well as an estimate of the probability and possible impact in case of a cyber incident. As soon as the inventory and analysis phase is completed, the authority responsible for the project should start creating the strategy. Specialized working groups could be created to study specific topics or design different sections of the strategy. The working groups should follow the processes defined in the initiation phase and adjust them if necessary. The implementation phase is more important in the NCS cycle. For the strategy to be successful, this means

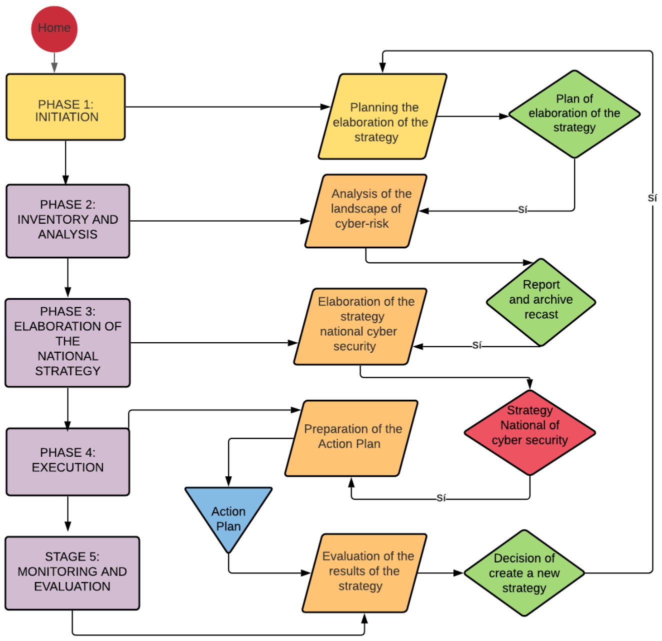

Home
PHASE 1: INITIATION
Planning the elaboration of the strategy
Plan of elaboration of the strategy
si
si
PHASE 2: INVENTORY AND ANALYSIS
Analysis of the landscape of cyber-risk
Report and archive recast
si
PHASE 3: ELABORATION OF THE NATIONAL STRATEGY
Elaboration of the strategy national cyber security
Strategy National of cyber security
PHASE 4: EXECUTION
Preparation of the Action Plan
si
Action Plan
STAGE 5: MONITORING AND EVALUATION
Evaluation of the results of the strategy
Decision of create a new strategy

Fig. 12. Strategic Cybersecurity Flowchart

that it is important to approach implementation in a structured and detailed manner with regard to appropriate human and financial resources and thus find a focus that should be seen as part of its development. During the follow-up phase, they should make sure that the strategy is implemented according to their action plan so that they do not have any problems in the future.During the assessment phase, the government and the relevant stakeholder should determine whether the strategy remains relevant to the evolving risks, whether it continues to meet the government's objectives, and whether adjustments are needed. In addition to assessing progress according to agreed metrics, it is important to periodically evaluate the results and compare them with the established objectives in order to manage the information well. In this evaluation it is important to understand if the objectives of the strategy are being met or if other actions need to be taken. As part of this process, the general risk context must also be reviewed periodically to see if any external changes have been made that could affect the results of the strategy, so by managing information well with up-to-date data and applying strategies to prevent cyber-attacks from cyber-criminals we can in one way or another prevent and combat the risks that may arise.

IV. Conclusions and Future Work

Our research article concludes after a thorough analysis and comparison of the different national strategies, thus explaining in detail the analysis of national strategies in cyber security and at the same time comparing cyber security strategies. In addition, a design has been made so that we are proposing a strategy design to combat cyber crime and through it help users to be able to prevent the different attacks by cyber crime, After analyzing the article and making the comparison it was shown that the most attacked country worldwide is Russia and in view of this a design has been made in which allows different countries to minimize the risks of different attacks that may occur in the world therefore our research work has been limited to make an analysis and design of national cybersecurity strategies. The article as a future work could expand more research implemented some computer systems that is to say that it allows to detect all these incidences mentioned that exist in cyber security worldwide.

References

[1] B. Collier, S. Horgan, R. Jones, and L. Shepherd, "The implications of the covid-19 pandemic for cybercrime policing in scotland: A rapid review of the evidence and future considerations," 2020.

[2] P. Chapman, "Are your it staff ready for the pandemic-driven insider threat?" *Network Security*, vol. 2020, no. 4, pp. 8–11, 2020.

[3] N. Shafqat and A. Masood, "Comparative analysis of various national cybersecurity strategies," *International Journal of Computer Science and Information Security*, vol. 14, no. 1, p. 129, 2016.

[4] V. Ibarra and M. ’o. n. Nieves, "International security determined by an online world: the state facing the challenge ’ i of terrorism and cybersecurity," in *VIII Congress of International Relations (La Plata, 2016)*, 2016.

[5] N. Shafqat and A. Masood, "Comparative analysis of various national cybersecurity strategies," *International Journal of Computer Science and Information Security*, vol. 14, no. 1, p. 129, 2016.

[6] F. Kolini and L. Janczewski, "Cluster and topic modeling: A new approach to national cyber security strategy analysis," in *Asia Pacific Conference on Information Systems (PACIS)*. Association of information systems, 2017.

[7] M. CARR, "Public–private partnerships in national cyber-security strategies," *International Affairs*, vol. 92, no. 1, pp. 43–62, 01 2016.

[8] K. Renaud, S. Flowerday, M. Warkentin, P. Cockshott, and C. Orgeron, "Is the responsibilization of the cyber security risk reasonable and judicious?" *computers & security*, vol. 78, pp. 198–211, 2018.

[9] S. A. ALOMARI, S. AL SALAIMEH, E. AL JARRAH, and M. S. ALZBOON, "Enhanced logistics information service systems performance: Using theoretical model and cybernetics' principles."

[10] "Elements of the national cybersecurity strategy for developing countries," *Magazine of the National Institute of Cybersecurity*, vol. 1, no. 3, pp. 9–19, 2015.

[11] M. F. Molina-Miranda, "Análisis de riesgos de centro de datos basado en la herramienta pilar de magerit," *Espirales revista multidisciplinaria de investigación*, vol. 1, no. 11, 2017.

[12] "Eeu and nato cybersecurity strategies and national cybersecurity strategies: a comparative analysis," *security journal*, vol. 30, no. 4, pp. 1151–1168, 2017.

[13] "Elementos de la estrategia nacional de ciberseguridad para países en desarrollo," *Revista del Instituto Nacional de Ciberseguridad*, vol. 1, no. 3, pp. 9–19, 2015.

[14] C. Haddad and C. Binder, "Governing through cyber-security: national policy strategies, globalized (in-) security and sociotechnical visions of the digital society,"

Österreichische Zeitschrift Für Soziologie, vol. 44, no. 1, pp. 115–134, 2019.

[15] D. Štitilis, P. Pakutinskas, M. Laurinaitis, and I. M.-v. de Castel, "A model for the national cyber security strategy. the lithuanian case." *Journal of Security & Sustainability Issues*, vol. 6, no. 3, 2017.

[16] S. J. Shackelford and A. Kastelic, "Toward a state-centric cyber peace: analyzing the role of national cybersecurity strategies in enhancing global cybersecurity," *NYUJ Legis. & Pub. Pol'y*, vol. 18, p. 895, 2015.

[17] F. Kolini and L. Janczewski, "Clustering and topic modelling: A new approach for analysis of national cyber security strategies," in *Pacific Asia Conference on Information Systems (PACIS)*. Association For Information Systems, 2017.

[18] I. Lütkebohle, "Directive (eu) 2016/1148 of the european parliament and of the council of 6 july 2016 concerning measures for a high common level of security of network and information systems across the union," 2016.

[19] D. Štitilis, P. Pakutinskas, and I. Malinauskaitė, "Eu and nato cybersecurity strategies and national cyber security strategies: a comparative analysis," *Security Journal*, vol. 30, no. 4, pp. 1151–1168, 2017.

[20] S. Abraham and S. Nair, "Comparative analysis and patch optimization using the cyber security analytics framework," *The Journal of Defense Modeling and Simulation*, vol. 15, no. 2, pp. 161–180, 2018.

[21] R. Sabillon, V. Cavaller, and J. Cano, "National cyber security strategies: global trends in cyberspace," *International Journal of Computer Science and Software Engineering*, vol. 5, no. 5, p. 67, 2016.

[22] S. Sengan, V. Subramaniyaswamy, S. K. Nair, V. Indragandhi, J. Manikandan, and L. Ravi, "Enhancing cyber–physical systems with hybrid smart city cyber security architecture for secure public data-smart network," *Future Generation Computer Systems*, vol. 112, pp. 724–737, 2020.

[23] L. P. Muller, "Cyber security capacity building in developing countries: challenges and opportunities," 2015.

[24] S. Enescu *et al.*, "A comparative study on european cyber security strategies," *Redefining Community in Intercultural Context*, vol. 9, no. 1, pp. 277–282, 2020.

[25] S. Sengan, V. Subramaniyaswamy, S. K. Nair, V. Indragandhi, J. Manikandan, and L. Ravi, "Enhancing cyber–physical systems with hybrid smart city cyber security architecture for secure public data-smart network," *Future Generation Computer Systems*, vol. 112, pp. 724–737, 2020.

[26] A. E. Shastitko, N. S. Pavlova, N. V. Kashchenko *et al.*, "Antitrust regulation of product ecosystems: The case study of kaspersky lab.–apple inc," *Upravlenets*, vol. 11, no. 4, pp. 29–42, 2020.

A Big Data Framework for Satellite Images Processing using Apache Hadoop and RasterFrames: A Case Study of Surface Water Extraction in Phu Tho, Viet Nam

Dung Nguyen[1], Hong Anh Le[2]
Faculty of Information Technology
Hanoi University of Mining and Geology
18 Pho Vien, Bac Tu Liem, Ha Noi, Viet Nam

***Abstract*—Earth data, collected from many sources such as remote sensing imagery, social media, and sensors, are growing tremendously. Among them, satellite imagery which play an important roles for monitoring environment and natural changes are increased exponentially in term of both volume and speed. This paper introduces an approach to managing and analyzing such data sources based on Apache Hadoop and RasterFrames. First, it presents the architecture and the general flow of the proposed distributed framework. Based on this, we can implement and perform efficient computations on a big data in parallel without moving data to the center computer which might lead to network congestion. Finally, the paper presents a case study that analyzes the water surface of a Vietnam region using the proposed platform.**

***Keywords*—Satellite imagery; big data; water surface; Apache Hadoop; RasterFrames**

I. Introduction

Data generated have been increased exponentially in recent years. The amount of data in recent two years is equal the to amount data which has been generated before. Data format also varies from structured to unstructured types. The format includes database in relational database systems, while the later might be video, images, or log files. Hence, 'big data' term appeared with three Vs such as *volume*, *velocity*, and *variety*. *Volume* refers to the amount of data generated, *Velocity* mentions the generation speed of the data, and *Variety* indicates the diversity of the data source.

One of most frequent generated sources is from earth observation. The big earth data consists of all data related to the earth including sensors, satellite images. Among them, satellite imaginary volume amount grows rapidly with advancement of technologies. For instance, in 2019, Sentinel (Sentinel-1, Sentinel-2, Sentinel-3), Landsat-7, Landsat-8, MODIS produce around 5 PB data [1], [2], [3]. For this reason, some solutions are proposed for building the platform for storing, managing, and processing EO data such as Google Earth Engine (GEE) [4], Sentinel Hub [5], Open Data Cube (ODC) [6], OpenEO [7], etc.. GEE is a platform that provides petabytes of satellite imagery and large-scale applicability for analysis. Sentinel Hub is an cloud-based engine for processing multi petabytes of satellite data. It allows users easily browse, visualize, and analyze Earth observation imagery. ODC is an open source project of geospatial data management and analysis. It provides facilities to work with raster data using Python libraries and PostgreSQL. These platform provide facilities to store and process satellite images both in commercial and open source solutions. In fact, however, there are some restricted policies of the organization or government that do not allow to store EO data in the cloud for example high-resolution satellite imagery from military areas or restricted areas.

Analyzing these data sources plays an important role in monitoring natural environment changes for instance land cover, surface water, body water, etc. Among them, surface water information is significant factor indicating the urbanization of an area and the management of water resources in the developing countries such as VietNam. Remote sensing and GIS techniques have been exploited to effectively analyze these changes. Recently, many research work proposed to use deep learning techniques and satellite imagery for surface water extraction and detection [8], [9], [10]. It, however, raises one issue regarding the huge volume of data to store and processing in a long time. For this reason, a simple and low-cost solution for this problem is desirable.

In this paper, we propose a framework for processing big satellite imagery data based on HDFS and Rasterframes. It allows to flexibly store satellite images in the distributed manner and perform manipulation on the data in parallel without moving data to the center. The contribution of this paper are (i) proposing a salable, open source, and low-cost framework for big satellite imaginary processing ; (ii) implementing and performing surface water analysis using sentinel-2 images for PhuTho area, a Northern province of VietNam.

The paper is structured as followed. Section II briefly give the introduction of Apache Hadoop platform and RasterFrames which are used as the basis of the proposed framework. Section III summarizes the related work. In the next section, the proposed framework is introduced in detail with its architecture and main flows. A case study of surface water extraction based on this framework is given in Section V-A. Finally, Section VI concludes the paper and presents the future work.

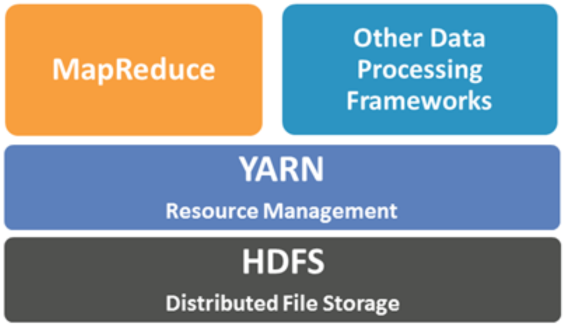


Fig. 1. Hadoop V2.0 Architecture

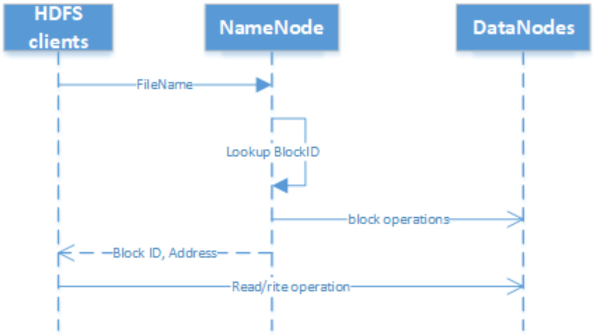


Fig. 2. Sequence Diagram of HDFS Operations

II. Background

A. Apache Hadoop

Apache Hadoop is an open source framework provides availability for distributed processing large data sets across single to thousands of nodes[11]. Hadoop was created by Nutch and a part of Apache Lunce. After implementing NDFS and MapReduce, in 2008, it became an independent one and rapidly reached the Apapche projects top-level in 2010. The basic archirecture of Apache Hadoop V2.0 is illustrated in Fig. 1. It has three main components including HDFS, MapReduce, and YARN.

- HDFS is designed for distributed file storage
- MapReduce provides features for parallel processing
- YARN stands for Yet Another Resources Negotiation that consists of ResourceManager and NodeManager. The former allocates resources between all application, the latter monitors the resource usage of the containers and report to ResourceManager.

Our proposed system utilizes HDFS as storage system, hence, this section introduces HDFS in detail. Files in HDFS are stored in independent block-size units that are much bigger than normal disk block size. A HDFS cluster has two different types of node namely namenode (or master node) and datanodes (worker nodes). NameNode manages namespace and block mapping to DataNodes, while data block units are stored in DataNodes. HDFS is designed to store data in large scale with reliability, therefore it makes a number of replication when files are ingested into the HDFS system. This factor is set to three as the default value. The HDFS reading and writing process of clients are illustrated in Fig. 2 that can be summarized as following steps

- HDFS clients send request to NameNode with a file name.
- NameNode looks up in its managed name space for data blocks ID of the requested file and returns to clients.
- NameNode frequently check DataNode with heartbeat.
- NameNode returns block ID and address to HDFS clients.
- Clients use a given block ID and perform reading and writing operations.

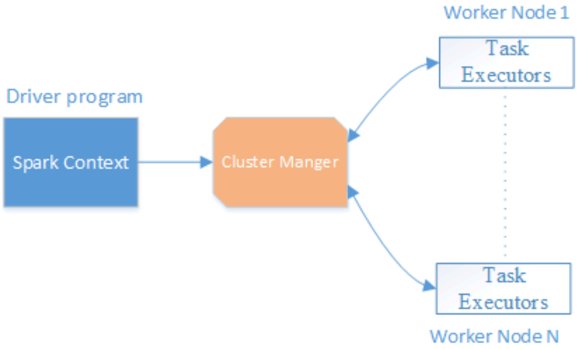


Fig. 3. Spark Applications Architecture

B. Apache Spark

Apache Spark is an analytics engine for large-scale data processing that supports high-level programming API in various languages and tools for querying, graph processing, and machine learning. It was started in 2009 in University of California, Berkeley RAD Lab and became an ASF top-level project in 2014. Spark performs 100 times faster than Hadoop Map Reduce for big data processing and can be deployed thourgh several other frameworks such as Mesos, Yarn, or standalone cluster. Fig. 3 depicts the basic architecture of Spark applications. The driver program in the Master Node first creates *Spark Context* that split jobs into tasks. Spark Context works with the *Cluster Manager* to mangage and distribute tasks over the clusters. Worder Nodes execute tasks and return resutls to Spark Context.

In order to program in Spark environment, this section introduces some basic concepts such as Resilient Distributed Dataset (RDD), DataFrame, and Dataset. RDD is immutablity data object that has various types and are stored in memory. DataFrame is semilar to relational database tables that is a distributed collection of data. DataFrame constructed from RDDs, tables in Hive, or external databases is conceptually equivalent to Pandas data frames. Listing 1 illustrate a Spark DataFrame object is created from a Parquet file.

Listing 1: Create Spark DataFrame from Parquet file

```
from pyspark.sql import *
parquetDF = spark.read.parquet("example.parquet")
```

C. Raster Frames

RasterFrames, a part of LocationTech project, provides functionalities to access earth observation data, cloud computing, and data analysis[12]. It allows to load and analyze raster data in DataFrames, to perform spatial queries and operator, to integrate with Apache Spark with powerful features in processing big data and machine learning library. RasterFrames has capability to read various types of raster data including GeoTIFF, JP2000, MRF, HDF via many protocols such as HTTP, FTP, HDFS, S3. Besides, it also can work with multiple spatial vector data types (point, line, polygon) from many data sources such as GeoJSON, WKT/WKB. It also supports for preserving geometry column while converting GeoDataFrame to SparkFrame. The Listing 2 shows the code for reading a GeoJSON file and convert multipolygon data in form of GeoDataFrame to SparkFrame. RasterFrames introduces a new type of data named *Tile* which is subset of a scene. A *Tile* is usually a two-dimensional array and square.

Listing 2: Reading GeoJSON and converting data with RasterFrames

```
from pyspark import SparkFiles
import geopandas
from shapely.geometry import MultiPolygon

data_uri = ’http://<uri>/example.geojson’
spark.sparkContext.addFile(data_uri)
df = spark.read.geojson(SparkFiles.get(’example’))

gdf = geopandas.read_file(data_uri)
gdf.geometry = gdf.geometry.apply(_multipolygons)
df2 = spark.createDataFrame(gdf)
```

In addition to features of handling with raster and vector data as well as sptial operator, RasterFrames also provides ability to train and predict with Apache Spark Machine Learning library(Spark MLLib) [13] that makes ML practical and easy. It consists of serveral libraries including ML algorithms (classification, regression, clustering, etc..), featurization, pipelines, persisence, and other ultilities. RasterFrames is built on top of Apache Spark, hence our proposed framework can be easily scaled up to run on cluster and cloud with the prorotyped solution developed on personal laptop.

III. Related Work

Gomes *et al.* [14] made a great review of seven platforms for big earth observation data management and analysis. Among them, ODC and SEPAN are open source ones which are similar to our proposed framework. The difference is that our proposed framework uses Hadoop and RasterFrames which can start with simple model but it can be extended later.

Rajak *et al.* [15] proposed a general framework for storing and processing high resolution satellite images using Apache Hadoop. Their work used MapReduce to perform computation and used HBase to store processed images. Our work is different and more concrete because we utilize RasterFrames with more powerfull features to analyze earth observation data.

Haifa Tamiminia*et al.* [16] made a systematic review of 349 articles that utilize Google Earth Engine for various big geodata applications. They showed the advantages of using GEE for big data processing with rich support features such as cloud storage, ML, DL algorithms.

Eken and Sayar [17] proposed a distributed big data framework for large scale raster mosaic images processing based on MapReduce programming model. The proposed framework consists of five steps including pre-processing, identifying mosaics, polygon coverage, stitching vectorised images, and object extraction and analysis.

Yanbo Huang *et al.* [18] presented a four-layer-twelve-level satellte remote sensing data management for agricultural remote sensing data including high-resolution satellites, UAV images, etc..

Huang *et al.* [19] introduced a scalable, efficient for heterogeneous remote sensing big data and management in distributed environment. The authors also used HDFS and HBase as storage layer. To query metadata contents, they constructed the index based on Elasticsearch.

Sedona *et al.* [20] proposed to use High Performance Computing machines to compute classification algorithms in parallel with CNNs.

Recently, Dalton Dunga [21] proposed RESFlow framework that includes several modular components such as clustering and embedding, image-bucket assignment, image gallery, model gallery, accelerated inference, application space,and image analytic. The framework provides parallel computing using Spark.

Arushi Patni [22] *et al.* introduced a model for satellite mages classification based on HDFS and Deep Neural Network. The result, however, just presented the architectural model with processing flow. It did not show any specific framework for processing satellite data and give any concrete example for their proposed model.

Hai Lan *et al.* [23] developed a Spark-based large-scale for processing remote sensing images. The authors used GAL to extract raster data and to transform to Spark RDDs. They proposed algorithms to caculate NDVI and NDWI values for classification and changing detection. The research also performed theree experiments with Landsat 8 and MODIS datasets on Amazon S3 and GEE.

Nguyen *et al.* [24] developed an end-to-end analysis pipeline using Caffe deep learning library along with two distributed platforms such as Spark and Dark.

Sukanta Roy*et al.* [25] also proposed to use Apache Hadoop as a solution to store remote sensing data. They implmented MapReduce programs for preprocessing Synthetic Aperture Radar and Multispectural data.

In comparision to these works, our approach is different because we use HDFS as infrastrucre along with Apache Spark and RasterFrames to process and analyze raster data at scale.

IV. Big Data Framework for Satellite Images Processing

Firstly, the general approach for processing big satellite imaginary data is presented, we then introduce the surface water extraction following the proposed approach for illustration purpose.

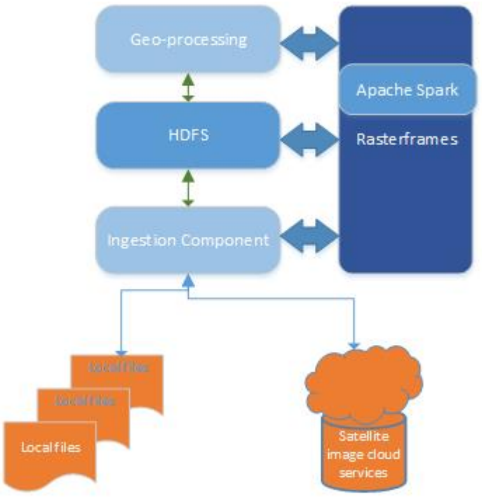


Fig. 4. Framework Architecture

A. Framework Architecture

Figure 4 depicts the architecture of for big satellite imagery processing framework. It based on HDFS and RasterFrame consists of two main layers that are describe as follows.

- The ingestion component: retrieves satellite imaginary data from local files or many cloud sources including Amazon S3, Sentinel-hub, etc.. This feature is implemented by using catalog and raster readers in RasterFrame. The *Ingestion* component ingests data and stores in HDFS clusters.
- The geo-processing component: proceeds and analyzes the raster data from HDFS or external sources. This component also reuse machine learning models with Spark ML Pipelines. These facilities are provided with RasterFrame.

B. Processing Workflow

Based on the proposed architecture, this section introduces the workflow of two main components: *Ingestion* and *Geoprocessing*.

Fig. 5 illustrates the flow of Ingestion. The input of the process is eitheir local raster files, external catalog in cloud-based services, or built-in catalogs supported by RasterFrames. These data are loaded in parallel into Spark DataFrames that can handle big data files. After loading to memories, if users are able process the in-memory data by invoking geoprocessing component in case of neccesity. Finally. it allows to store in HDFS with original file format or Apache Parquet files.

Fig. 6 depicts the general flow of information mining from imaginery. At the first step, it loads data stored in HDFS or external catalog from AWS Public Data Set (PDS). Next two steps are data preparation and data quality enhancement. In train and predict step, we choose the appropriated algorithms or models that provided SparkMLLib for each specific task. In case of supervised learninig models, label data is required and joined with the prepared data. After prediction, in the last step, we evaluate the model, visualize the results, or serialize to the proposed HDFS system.

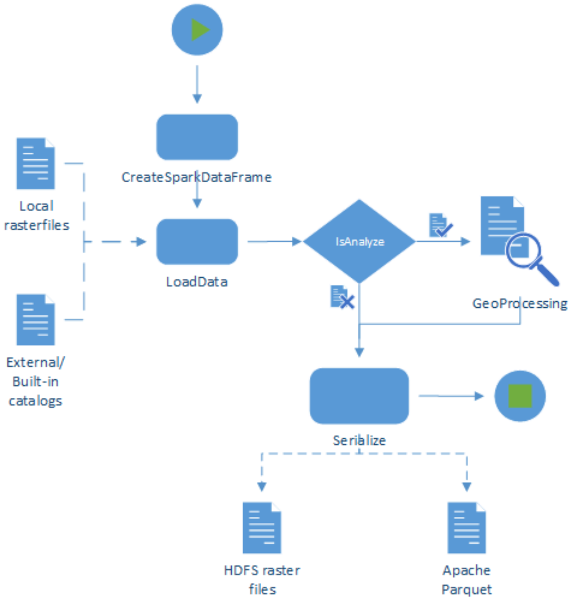


Fig. 5. Ingestion Work Flow

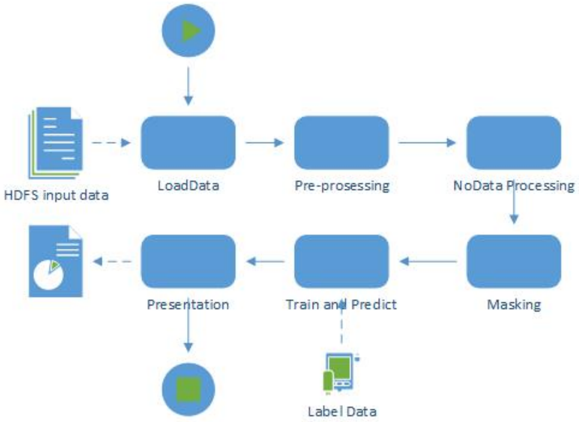


Fig. 6. Geoprocessing Work Flow

V. A Case Study: Surface Water Extraction in Phu Tho Province, VietNam

This section presents a case study of surface water extraction with Sentinel-2 images following the approach introduced in section IV-A. First, we introduce the region and used data. The data sources are Sentinel-2 images of the study area which are stored as local files in single computer nodes. These images are stored into HDFS clusters for processing later using ingestion layer. For surface water extraction, we follow the proposed framework and use a supervised machine learning technique which is built inside RasterFrames.

A. Study Area and Used Data

Phu Tho is a mountainous province in North located at the centre of Vietnam's western Northeast. Phu Tho has 13 district, city and township with five large rivers including Chay, Hong, Da, Lo, and Bua rivers and many branches which provide more than 2,500 ha surface water are for agriculture development.

It also has various group of residents and variety of living culture. Phu Tho has been a industrial province with theree regions Viet Tr, Bai Bang - Lam Thao, and Thanh - Ha Hoa. Currently, surface water quality in PhuTho has been quickly polluted at the alarm level caused by many factors such as industrialization, urbanization, and living styles.

The case study uses Sentinel-2 images as the data source which were collected in range from 2019 to 2020 and pushed into HDFS using *Ingestion* component. The Sentinel-2 mission consists of two satellite systems including Sentinel-2A and Sentinel-2B, which were launched on 23 June 2015 and 07 March 2017 respectively. These satellites provide a global coverage of the Earth's land surface every five days.

B. Main Process

Figure 7 depicts the detailed work flow of extraction process that consists of the following steps

- Load data from HDFS: After the local Sentinel-2 images are ingested to HDFS, the selected are data are loaded into data frames.

- Remove low quality data: It is the masking operation to set values to missing data or cloudy cells. The SCL (scene classification) band is used to check if a cell contains no data, saturated, cloud probability, or vegetation i.e., the cell value is in range of 0,1,4,8,9,10. It can be implemented by invoking function *rf_local_is_in[0,1,4,8,9,10]*.

- Create label data: Creating polygon regions in form of GeoJSON to indicate surface water areas. This small DataFrame is joined later with raster DataFrame loaded from HDFS.

- Create SparkML pipelines: This step utilizes Apache Spark ML pipelines to create a pipeline which consists of four stages including tile exploding, data filtering, assembler and classifier. For surface water extraction, we use Decision Tree and Random Forest algorithms to evaluate. Listing V-B shows the snippet of Python code to construct the pipeline with Random Forest algorithm.

Listing 3: Create SparkML pipelines

```
classifier = RandomForestClassifier() \
  .setLabelCol('label') \
  .setFeaturesCol(assembler.getOutputCol())
pipeline = Pipeline() \
  .setStages([tileExploder, dataFilter, \
        assembler, classifier])
```

- Training and Prediction: Execute the pipeline created in the previous step and make model prediction applying *transformation* method on the pipeline as follows.

```
 model = pipeline.fit(model_input)
 pre = model.transform(df_mask) \
.drop(assembler.getOutputCol()).cache()
```

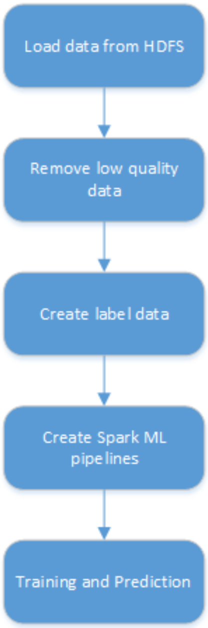


Fig. 7. Surface Water Extraction Flow

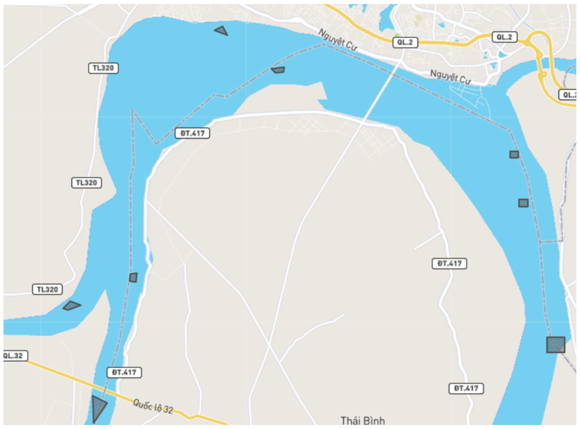


Fig. 8. Create Label Data for Surface Water

C. Implementation and Results

As mentioned in the previous section, we implement the surface water in Phu Tho province. After reading raster data of the study area from HDFS, we create a small set of data including 26 sample label of surface water in Phu Tho across the several Northern provinces in VietNam. These data can be created by add features in geojson.io illustrated in Fig. 8.

Fig. 9 shows the three extracted scenes and the prediction results of the study region that use DecisionTree algorithms and satellite images collected in April 2020.

We also can evaluate the correctness of the models by using the following the snippets (Listing 4) to show the accuracy and confusion matrix of the model.

Listing 4: Evaluate the model

```
 model_eval = MulticlassClassificationEvaluator(
 prediction=classifier.getPredictionCol(),
```

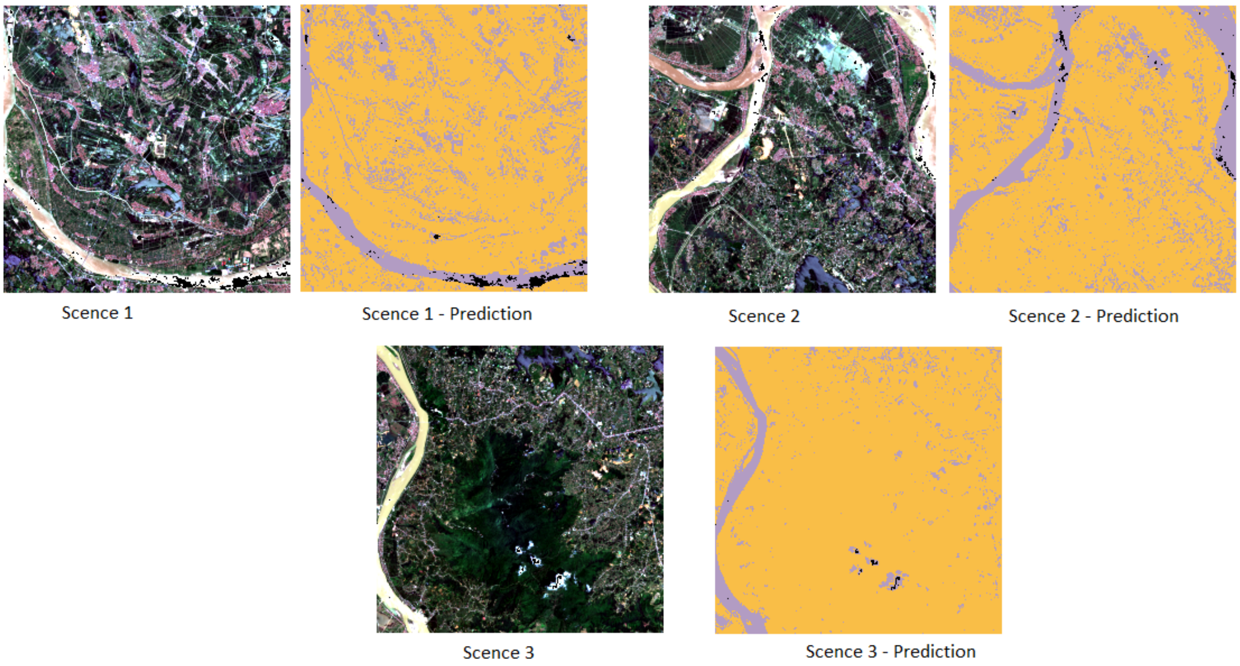


Fig. 9. Prediction results of extracted scenes

```
label=classifier.getLabelCol(),
metricName='accuracy')
accuracy = model_eval.evaluate(prediction_ret)
confusion_matrix = prediction_df.groupBy
(classifier.getPredictionCol()) \
.pivot(classifier.getLabelCol()) \
.count() \
.sort(classifier.getPredictionCol())
```

The prediction accuracy of DecisionTree model is slightly better with 0.959 while RandomForest gives the result at 0.954 those confusion matrixes are shown in Table I and Table II respectively.

TABLE I. Confusion matrix with Decision Tree algorithm

prediction	Water	Non-Water
Water	354	17
Non-Water	46	

TABLE II. Confusion matrix with Random Forest algorithm

prediction	Water	Non-Water
Water	353	18
Non-Water	1	45

VI. Conclusion and Future work

In this paper, we propose a framework for satellite images processing in large scale. It based on the open source platforms such as Apache Hadoop and RasterFrames consists of two major components including storage layer with HDFS architecture and processing layer that combines Spark and RasterFrames. The proposed framework can be built on the commodity hardware and has the low cost for deployment. The paper also implements a surface water extraction with Sentinel-2 images in PhuTho province Viet Nam with high accuracy for illustration purpose. The case study shows the flexibility of the proposed framework. We intend to develop the framework with more powerful features such as deep learning libraries for object segmentation and detection.

Acknowledgment

This research is funded by Hanoi University of Mining and Geology, Vietnam (Project No. T19-29).

References

[1] J. Salazar Loor and P. Fdez-Arroyabe, *Aerial and Satellite Imagery and Big Data: Blending Old Technologies with New Trends*. Cham: Springer International Publishing, 2019, pp. 39–59.

[2] H.-D. Guo, L. Zhang, and L.-W. Zhu, "Earth observation big data for climate change research," *Advances in Climate Change Research*, vol. 6, no. 2, pp. 108 – 117, 2015, special issue on advances in Future Earth research. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1674927815000519

[3] Y. Ma, H. Wu, L. Wang, B. Huang, R. Ranjan, A. Zomaya, and W. Jie, "Remote sensing big data computing: Challenges and opportunities," *Future Generation Computer Systems*, vol. 51, pp. 47 – 60, 2015, special Section: A Note on New Trends in Data-Aware Scheduling and Resource Provisioning in Modern HPC Systems. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X14002234

[4] (2020) Google earth engine. https://earthengine.google.com/.

[5] (2020) Sentinel Hub. https://sentinel-hub.com/.

[6] (2020) Open Data Cube. https://opendatacube.org/.

[7] (2020) OpenEO. https://openeo.org/.

[8] T. D. Acharya, A. Subedi, H. Huang, and D. H. Lee, "Classification of surface water using machine learning methods from landsat data in nepal," *Proceedings*, vol. 4, no. 1, 2019. [Online]. Available: https://www.mdpi.com/2504-3900/4/1/43

[9] W. Huang, B. DeVries, C. Huang, M. Lang, J. Jones, I. Creed, and M. Carroll, "Automated extraction of surface water extent from sentinel-1 data," *Remote Sensing*, vol. 10, no. 5, p. 797, May 2018. [Online]. Available: http://dx.doi.org/10.3390/rs10050797

[10] A. Sekertekin, S. Y. Cicekli, and N. Arslan, "Index-based identification of surface water resources using sentinel-2 satellite imagery," in *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2018, pp. 1–5.

[11] (2020) Apache Hadoop. https://hadoop.apache.org.

[12] (2020) Rasterframes. https://rasterframes.io/.

[13] (2020) Apache Spark. https://spark.apache.org/.

[14] V. C. F. Gomes, G. R. Queiroz, and K. R. Ferreira, "An overview of platforms for big earth observation data management and analysis," *Remote Sensing*, vol. 12, no. 8, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/8/1253

[15] R. Rajak, D. Raveendran, M. C. Bh, and S. S. Medasani, "High resolution satellite image processing using hadoop framework," in *2015 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, 2015, pp. 16–21.

[16] H. Tamiminia, B. Salehi, M. Mahdianpari, L. Quackenbush, S. Adeli, and B. Brisco, "Google earth engine for geo-big data applications: A meta-analysis and systematic review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 164, pp. 152 – 170, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271620300927

[17] S. Eken and A. Sayar, "A mapreduce based big-data framework for object extraction from mosaic satellite images," 2018.

[18] Y. Huang, Z. xin CHEN, T. YU, X. zhi HUANG, and X. fa GU, "Agricultural remote sensing big data: Management and applications," *Journal of Integrative Agriculture*, vol. 17, no. 9, pp. 1915 – 1931, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2095311917618598

[19] X. Huang, L. Wang, J. Yan, Z. Deng, S. Wang, and Y. Ma, "Towards building a distributed data management architecture to integrate multi-sources remote sensing big data," in *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 2018, pp. 83–90.

[20] R. Sedona, G. Cavallaro, J. Jitsev, A. Strube, M. Riedel, and J. A. Benediktsson, "Remote sensing big data classification with high performance distributed deep learning," *Remote Sensing*, vol. 11, no. 24, 2019.

[21] D. Lunga, J. Gerrand, L. Yang, C. Layton, and R. Stewart, "Apache spark accelerated deep learning inference for large scale satellite image analytics," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 271–283, 2020.

[22] A. Patni, K. Chandelkar, R. K. Chakrawarti, and A. Rajavat, "Classification of satellite images on hdfs using deep neural networks," in *International Conference on Advanced Computing Networking and Informatics*, R. Kamal, M. Henshaw, and P. S. Nair, Eds. Singapore: Springer Singapore, 2019, pp. 49–55.

[23] H. Lan, X. Zheng, and P. M. Torrens, "Spark sensing: A cloud computing framework to unfold processing efficiencies for large and multiscale remotely sensed data, with examples on landsat 8 and modis data," *Journal of Sensors*, vol. 2018, p. 2075057, Aug 2018. [Online]. Available: https://doi.org/10.1155/2018/2075057

[24] M. H. Nguyen, J. Li, D. Crawl, J. Block, and I. Altintas, "Scaling deep learning-based analysis of high-resolution satellite imagery with distributed processing," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 5437–5443.

[25] S. Roy, S. Gupta, and S. Omkar, "Case study on: Scalability of preprocessing procedure of remote sensing in hadoop," *Procedia Computer Science*, vol. 108, pp. 1672 – 1681, 2017, international Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050917305598

The Influence of Loss Function Usage at SIAMESE Network in Measuring Text Similarity

Suprapto[1], Joseph A. Polela[2]
Department of Computer Science and Electronics
Universitas Gadjah Mada
Yogyakarta, Indonesia

***Abstract*—In a text matching similarity task, a model takes two sequence of text as an input and predicts a category or scale value to show their relationship. A developed model is to measure the similarity - one of relationship between those two text. The model is SIAMESE network that implement two copies of same network of CNN, it takes $text\_1$ and $text\_2$ as the inputs respectively for two CNN networks. The output of each CNN network is features vector of the corresponding text input, both outputs are then fed by a loss function to calculate the value of loss (i.e. similarity). This research implemented two types of loss functions, i.e. Triplet loss and Contrastive loss. The usage purpose of these two types of loss functions was to see the influence toward the measurement results of similarity between two text being compared. The metrices used for this comparison are $precision$, $recall$, and F1-$score$. Based on the experimental results done on 1500 pairs of sentences, and varied on the epoch value starting from 10 until 200 with an increment of 10, showed the best result was for epoch value of 180 with $precision$ 0.8004, $recall$ 0.6780, and F1-$score$ 0.6713 for Triplet loss function; and epoch value of 160 with $precision$ 0.6463, $recall$ 0.6440, and F1-$score$ 0.6451 for Contrastive loss function gave the best performance. So that, the Triplet loss function gave better influence than Contrastive loss function in measuring similarity between two given sentences.**

***Keywords*—Sentence; similarity; triplet; contrastive; CNN; Siamese; dataframe**

I. Introduction

The very fast growth of information nowdays causes a particular problem, such as an overwhelming of information [21]. It is very likely among those collections of huge of information found some similar ones, so that, they can be grouped into several classes based on their similarity. In order to overcome the overwhelming of information, each class will only be represented by a single information. Obviously, to identify the similar information requires a process called similarity measurement. A text similarity measurements is one of text mining approach that capable of coping with the information overwhelming. This process begins with finding similar word for sentece, then paragraph, and finally document [6]. Text similarity approach will ease people to find relevance information. It has a great support in successness for text mining operations such as, searching and information retrieval (IR), text classification, information extraction (IE), document clustering [8], sentiment analysis [4] [10] [16][3] [13], machine translation, text summarization, and natural language processing (NLP). Text similarity measurement may be done by comparing text - text matching. In text comparison tasks, a model takes two texts as inputs and predicts a category or a scale value indicates the relationship between those two texts. A big number of varieties of tasks such as, natural language inference [2] [11], paraphrase identification [17], answer selection [19] could be consider as special form of text matching problems. Recently, deep neural network is the most popular choice for text mining. Semantic alignment and comparison of two sequence of texts are the key of neural text matching. Most of previous deep neural network contain single inter-sequence alignment layer. In order to make the alignment process fully used, model must take many external syntaxtical features or aligment as additional inputs at alignment layer [5] [7], adopt a complex alignment mechanism [17], or build a big number of post-process layers to analyze alignment results [7].

This research proposed two models to compute similarity value between two texts using SIAMESE network in which each uses two different types of loss function - Triplet loss and Contrastive loss. The model consists of two copies of same CNN networks, where each recieves text (or sentence) as input. Subsequently, each CNN network results feature vector of each recieved text, and finally fed by loss function to compute the similirity value. Each model will be tested using the same dataset **Quora Question Pair similarity** taken from https://www.kaggle.com/c/quora-question-pairs. In order to see the influence of using these two types of loss function, the three metrics: $precision$, $recall$, and F1-$score$ will be computed.

II. Related Works

Deep neural network is very dominant in text matching area. While semantic alignment and comparison between two text sequences is the core of text matching [17]. The very beginning task explores encoding of each sequence individuallly into a vector and then bulds a neural network classifier over the two vectors. In this paradigm, recurrence [2], recursive [12], convolutional network [20] were used as sequence encoder. In this model, where encoding from a sequence independent with other sequence caused the last classifier had difficulty in modeling complex relations. Therefore, the subseqeunt tasks adopt framework matching aggregation to match two sequences at the lower level and aggregate results based on attention mechanism. DecomAtt used a simple form from attention for alignment and aggregate representations aligned by feed-foward network [15]. ESIM used a similar attention mechanism and implement bidirectional LSTMs as encoders and aggregators [5]. In order to improve model's performance, the researcher adopted three main paradigms. The first paradigm used syntacs that richer hand-writing features. HIM used syntactic parse tree [5]. The many usages of POS were found in previous tasks, some of them in [12]

[7]. The exact match with lemmatized tokens was reported as a powerfull binary features [7][12]. The second way was adding the complexity to alignment computation. BiMPM exploited an advanced multi-perspective matching operation [17], and MwAN implemented multi heterogeneous attention functions to calculate the alignment's results. The third way to improve model is by building heavy post-processing layers for alignment results. CAFE extracted additional indicators from allignment process using alignment factorization. DIIN adopted DenseNet as a deep convolutional feature extractor to filter information from alignment results [7]. The more effective models could be built when inter-sequence matching was allowed to be done more than once. CSRAN performed multi-level attention refinement with dense connections among many levels [12]. DRCN stacked encoding and alignment layers [12]. DRCN concatenated all previous alignment results and must've used autoencoder to cope with features space explotion. SAN exploited recurrent networks to combine many alignment results [14]. An architecture of deep based on new way relating contiguous blocks called by augmented residual connections to filter previous aligned information roling as inmportant features for text matching [18].

III. Methodology

This section presents the building of two Siamese networks each with **triplet loss** and **contrastive loss** functions.The training model for Siamese network with triplet loss function consists of three copies of same network of CNN, it takes $text\_1$, $text\_2$ and $text\_3$ as the inputs, while one with contrastive loss function consists of two copies only, it takes $text\_1$, and $text\_2$ as the inputs. However, the testing model for Siamese network both with triplet and contrastive loss function consist of two copies of same network of CNN, it takes $text\_1$, and $text\_2$ to be calculated their similarity. The dataset used to do training and testing was **Quora Question Pair similarity** taken from https://www.kaggle.com/c/quora-question-pairs. At the end, the three metrics: precision, recall, and F1-score were computed to see the influence of loss function's usage in each network.

A. Learning Model

Based on the objective of the research, the model was built with three main components: twin network, similarity function, and output layer.

1) **Twin Network**: most feature's extraction for text take place in this network. It has two copies of same networks, two networks share set of same weights. It capable of receiving two different inputs - two sequence of text. Each network in it is a convolutional text encoder. CNN network is used twice before performing backpropagation.
2) **Similarity function**: two outputs from twin netork representing features learned automatically from two compared text was fed by a layer. Subsequently, distance formula was used to compare the similarity between two text in *n*-dimensional space.
3) **Output layer**: last layer whose single neuron connecting *n* neurons from the results of previous similarity function. The role of this part of the model was to decide the probability of tested text to be a member the text used as reference. The probability was computed using **Sigmoid** function as in equation (1).

$$p = \sigma(\Sigma_j \sigma_j |h_j^1 - h_j^2|) \quad (1)$$

where h_j^1 and h_j^2 are values of j-th neuron of first twin network and second twin network respectively, σ_j is weight between neuron output and j-th neuron in similarity layer.

The structure of **Siamese** network is shown in Fig. 1 [23].

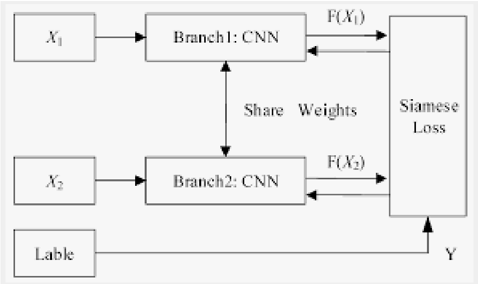


Fig. 1. The Structure of **Siamese** Network

The two types of loss function are implemented in the research, namely, **triplet** and **contrastive**. The aim of this implementation is to find the influence of loss function usage in similarity value computed between two text.

a. **Triplet Network**
A **triplet** network is comprised of three instances of the same feedforward network (with shared parameters). When fed with three samples, the network outputs two intermediate values - the L_2 distances between the embedded representation of two of its inputs from the representation of the third. The three inputs will be denoted as x, x^+ and x^-, and the embedded representation of the network as $Net(x)$. In words, this encodes the pair of distances between each of x^+ and x^-, against the reference x, i.e., $||Net(x) - Net(x^-)||_2$ and $||Net(x) - Net(x^+)||_2$. The distance between reference input x and positive input x^+ was minimized, on the other hand, the distance between reference input x and negative input x^- was maximized. The structure of **triplet** network is shown in Fig. 2 [9].

b. **Contrastive Network**
The **contrastive** loss function takes output from network for positive sample and computes the distance to an example from the same class and contrast it with distance to negative examples. In other word, the value of loss is low if positive samples are encoded to similar representations (closer), and negative examples are encoded to different representations (farer). The formula used to compute distance was **cosine similarity** distance as shown in equation (2).

$$\begin{aligned} Loss(x_p, x_q, y) = y * ||x_p - x_q||^2 + (1-y) * \\ max(0, m^2 - ||x_p - x_q||^2) \end{aligned} \quad (2)$$

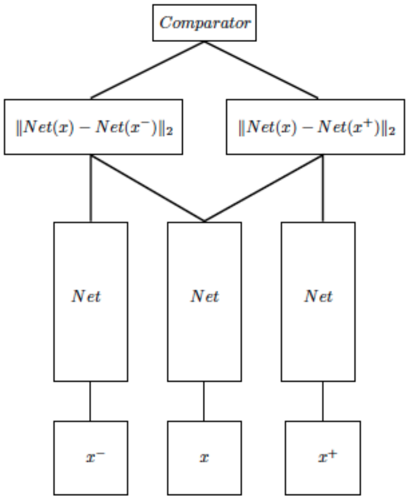


Fig. 2. The Structure of **Triplet** Network

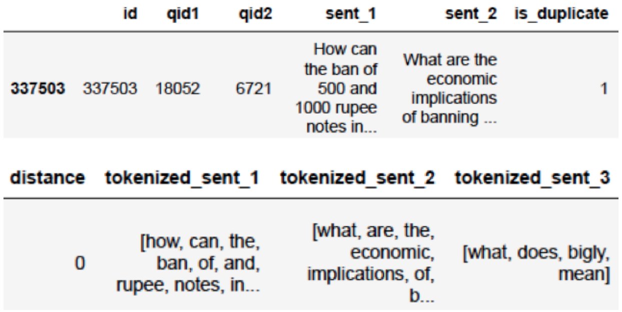
IV. Results and Discussion

As previously mentioned, the dataset used in the research was public. It was taken from https://www.kaggle.com/c/quora-question-pairs. The dataset was in CSV file containing 5000 pairs of Quora questions. The dataset was read to generate dataframe, and then the generated dataframe was devided into two: 70% for training and the other 30% for testing. The next process was building Word2Vec model using Gensim for embedding layer of deep learning. Then define function to transform sentence into vector containing index of Word2Vec's vocabularies. Another two important functions to be defined were **triplet** and **contrastive** loss, continued to define the CNN Siamese network with both **triplet** and **contrastive** loss functions, and then trained it.

A. Dataframe

The dataframe was generated in several steps: first applying simple process both to question 1 ($sent\_1$) and question 2 ($sent\_2$), to do this there was a function $simple\_process$ taken from https://radimrehurek.com/gensim/utils.html #gensim.utils.simple_preprocess for tokenization. This process was done fao all pairs of sentences in the dataset. From these all tokenized pairs of sentence, the number of tokens from the longest sentence was calculated based on the *mean* and *deviation standard*. This value was used to do padding, to made all tokenized pairs of sentences had the same length. In order to train the Siamese CNN model with triplet loss function, a negative sentence (the third tokenized sentence) must have been added to each pair of tokenized sententece (i.e. tokenized_sent_1 and tokenized_sent_2) for training dataframe. The format of training dataframe is shown in Fig. 3.

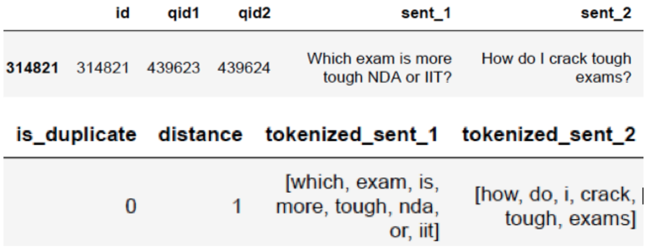
However, in the testing process for Siamese CNN model even with triplet loss function only need two tokenized sentences (tokenized_sent_1 and tokenized_sent_2). The dataframe format for testing was shown in Fig. **??**.



	id	qid1	qid2	sent_1	sent_2	is_duplicate
337503	337503	18052	6721	How can the ban of 500 and 1000 rupee notes in...	What are the economic implications of banning ...	1

distance	tokenized_sent_1	tokenized_sent_2	tokenized_sent_3
0	[how, can, the, ban, of, and, rupee, notes, in...	[what, are, the, economic, implications, of, b...	[what, does, bigly, mean]

Fig. 3. The Format of Training Dataframe



	id	qid1	qid2	sent_1	sent_2
314821	314821	439623	439624	Which exam is more tough NDA or IIT?	How do I crack tough exams?

is_duplicate	distance	tokenized_sent_1	tokenized_sent_2
0	1	[which, exam, is, more, tough, nda, or, iit]	[how, do, i, crack, tough, exams]

Fig. 4. The Format of Testing Dataframe

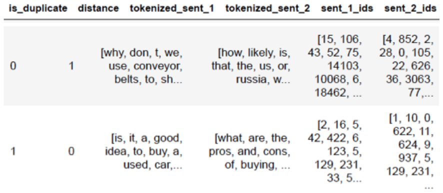
B. Word2Vec Model

Before building the CNN **Siamese** network, the Word2Vec model was built using Gensim for embedding layer in deep learning. It was built with CBOW (https://iksinc.online/tag/continuous-bag-of-words-cbow/) with 20 iterations, vector length of 100, and window size was 5. The Word2Vec was trained once using string of words (tokenozed sentence as a result of concatenation between tokenized of sentence 1 and tokenized of sentence 2), and saved it. During the process of training the model, there was a function required to convert a sentence into a vector containing Word2Vec vocabulary indices. For the model using **triplet** loss function, the converter changed three tokenized sentences (i.e. tokenized_sent_1, tokenized_sent_2, tokenized_sent_3) into three vectors containing indices of Word2Vec vocabularies (i.e. sent_1_ids, sent_2_ids, sent_3_ids), and two tokenized sentences (i.e. tokenized_sent_1, tokenized_sent_2) into two vectors containing indices of Word2Vec vocabulary (i.e. sent_1_ids, sent_2_ids) respectively in training and testing processes. However, for the model using **contrastive** loss function, the converter changed two tokenized sentences (i.e. tokenized_sent_1, tokenized_sent_2) into two vectors containing indices of Word2Vec vocabularies (i.e. sent_1_ids, sent_2_ids) in both training and testing processes. The sample of conversion results was shown in Fig. 5.

While the definition of the function that convert list of tokens to list of indices was shown in Fig. 6.

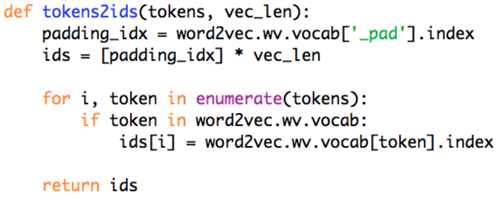
C. Triplet Loss

The **Triplet** loss function takes three inputs: baseline (an anchor sentence), positive (true sentence - one closest to an anchor), and negative (false sentence - one farest to an anchor). Therefore, the objective of this function is to minimize the



is_duplicate	distance	tokenized_sent_1	tokenized_sent_2	sent_1_ids	sent_2_ids
0	1	[why, don, t, we, use, conveyor, belts, to, sh...	[how, likely, is, that, the, us, or, russia, w...	[15, 106, 43, 52, 75, 14103, 10068, 6, 18462, ...	[4, 852, 2, 28, 0, 105, 22, 626, 36, 3063, 77,...
1	0	[is, it, a, good, idea, to, buy, a, used, car,...	[what, are, the, pros, and, cons, of, buying, ...	[2, 16, 5, 42, 422, 6, 123, 5, 129, 231, 33, 5...	[1, 10, 0, 622, 11, 624, 9, 937, 5, 129, 231, ...

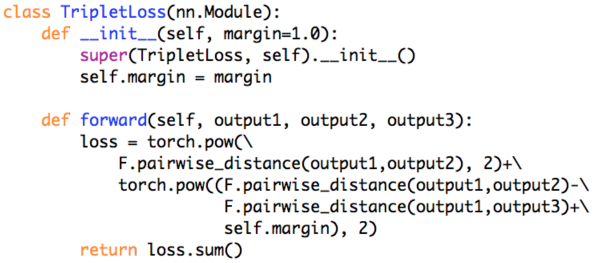
Fig. 5. The Conversion Result Sample



```
def tokens2ids(tokens, vec_len):
    padding_idx = word2vec.wv.vocab['_pad'].index
    ids = [padding_idx] * vec_len

    for i, token in enumerate(tokens):
        if token in word2vec.wv.vocab:
            ids[i] = word2vec.wv.vocab[token].index

    return ids
```

Fig. 6. The Definition of Converter Function

distance between an anchor sentence with a positive sentence, on the other hand, maximize the distance between an anchor sentence with a negative sentence. The implementation of the **triplet** function was shown in Fig. 7.



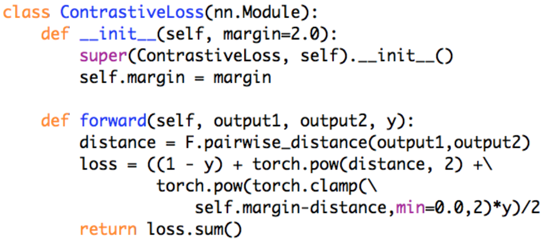
```
class TripletLoss(nn.Module):
    def __init__(self, margin=1.0):
        super(TripletLoss, self).__init__()
        self.margin = margin

    def forward(self, output1, output2, output3):
        loss = torch.pow(\
            F.pairwise_distance(output1,output2), 2)+\
            torch.pow((F.pairwise_distance(output1,output2)-\
                       F.pairwise_distance(output1,output3)+\
                       self.margin), 2)
        return loss.sum()
```

Fig. 7. The Implementation of Triplet Loss Function

D. Contrastive Loss

Unlike the **triplet**, the **contrastive** loss function only takes two inputs: positive sample and negative example. The objective is to contrast the distance between positive sample and an example from the same class with the distance between positive sample and negative example.

The implementation of the **triplet** function was shown in Fig. 8.



```
class ContrastiveLoss(nn.Module):
    def __init__(self, margin=2.0):
        super(ContrastiveLoss, self).__init__()
        self.margin = margin

    def forward(self, output1, output2, y):
        distance = F.pairwise_distance(output1,output2)
        loss = ((1 - y) + torch.pow(distance, 2) +\
                torch.pow(torch.clamp(\
                    self.margin-distance,min=0.0,2)*y)/2
        return loss.sum()
```

Fig. 8. The Implementation of Contrastive Loss Function

These two functions were only used during the training process to update the weights of network in order to converge; the **triplet** loss used two distances, and the **contrastive** loss only used one distance. While during the testing, no loss function needed, but the distance between two sentences. Because the model simply used the final weights resulted from training process.

E. CNN Siamese Network

The CNN **Siamese** network was built with the following specifications:

1) The embedding layer was the Word2Vec model that had already been trained.
2) Three 2-dimensional convolutional layers with configurable hyperparameters.
3) For every convolutional layer used $tanh$ as an activation function, dropout, and maxpool layer.

The architecture of CNN that was implemented in the CNN Siamese or just Siamese model was shown in Fig. 9 [22].

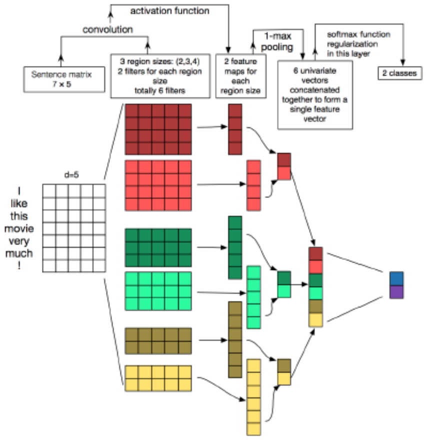
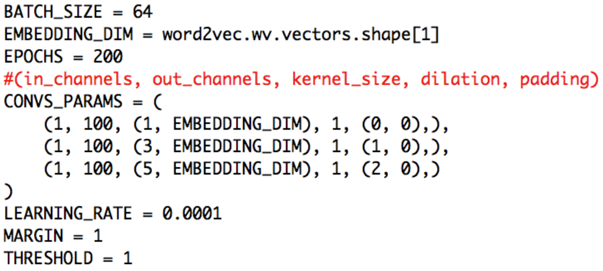

activation function
convolution
1-max pooling
softmax function regularization in this layer
Sentence matrix 7 x 5
3 region sizes: (2,3,4) 2 filters for each region size totally 6 filters
2 feature maps for each region size
6 univariate vectors concatenated together to form a single feature vector
2 classes
d=5
I
like
this
movie
very
much
!

Fig. 9. The Architecture of Implemented CNN

The hyperparameter of the CNN **Siamese** network was shown in Fig. 10.



```
BATCH_SIZE = 64
EMBEDDING_DIM = word2vec.wv.vectors.shape[1]
EPOCHS = 200
#(in_channels, out_channels, kernel_size, dilation, padding)
CONVS_PARAMS = (
    (1, 100, (1, EMBEDDING_DIM), 1, (0, 0),),
    (1, 100, (3, EMBEDDING_DIM), 1, (1, 0),),
    (1, 100, (5, EMBEDDING_DIM), 1, (2, 0),)
)
LEARNING_RATE = 0.0001
MARGIN = 1
THRESHOLD = 1
```

Fig. 10. The CNN **Siamese**'s Hyperparameter

The CNN **Siamese** model with **triplet** loss function consisting of three exact same models of CNN was trained using triple tokenized sentences (i.e. tokenized_sent_1, tokenized_sent_2, and tokenized_sent_3) from the dataframe training. The output of each CNN model was fed by **Triplet** loss function to calculate the loss value. On the other hand, the CNN **Siamese** model with **contrastive** loss function consisting of two exact same models of CNN was trained using pair of tokenized sentences (i.e., tokenized_sent_1, and tokenized_sent_2) from the dataframe training. The output of each CNN model was fed by **contrastive** loss function to calculate the loss value. Finally, both model of CNN **Siamese** with **triplet** and **contrastive** loss function respectively were tested by feeding two tokenized sentences to be calculated their similirity value. Both testing were conducted using the same number of 1500 pairs of sentences (Quora questions), and varied for the values of EPOCH starting from 10 until 200 with an increment of 10 to obtain the values of three metrics: precision, recall and F1-score. The values of three matrics for each CNN **Siamese** model were shown respectively in Table I and Table II.

TABLE I. THE PRECISIONS, RECALLS AND F1-SCORES OF THE CNN **SIAMESE** WITH **TRIPLET**

Epoch	Training Loss	Precision	Recall	F1-Score
10	64.3868	0.3762	0.6133	0.4663
20	19.4053	0.3762	0.6133	0.4663
30	8.6268	0.6277	0.6187	0.4896
40	4.9022	0.6469	0.6447	0.5753
50	3.2373	0.6762	0.6827	0.6602
60	2.3511	0.7073	0.7127	0.7074
70	1.8006	0.7254	0.7207	0.7224
80	1.4510	0.7433	0.7253	0.7288
90	1.2047	0.7526	0.7160	0.7194
100	1.0458	0.7520	0.6980	0.7003
110	0.9170	0.7562	0.6873	0.6880
120	0.8245	0.7746	0.6960	0.6956
130	0.7459	0.7774	0.6860	0.6838
140	0.6954	0.7834	0.6807	0.6768
150	0.6398	0.7892	0.6800	0.6752
160	0.6074	0.7885	0.6707	0.6642
170	0.5824	0.7907	0.6707	0.6639
180	0.5589	0.8004	0.6780	0.6713
190	0.5404	0.7973	0.6700	0.6621
200	0.5203	0.7986	0.6733	0.6659

TABLE II. THE PRECISIONS, RECALLS AND F1-SCORES OF THE CNN **SIAMESE** WITH **CONTRASTIVE**

Epoch	Training Loss	Precision	Recall	F1-Score
10	16.2760	0.3762	0.6133	0.4663
20	5.8481	0.3762	0.6133	0.4663
30	2.6012	0.3760	0.6127	0.4660
40	1.4245	0.6290	0.6193	0.4921
50	0.8705	0.6370	0.6333	0.5434
60	0.5853	0.5921	0.6200	0.5508
70	0.4222	0.6068	0.6287	0.5824
80	0.3151	0.6113	0.6313	0.5976
90	0.2358	0.6071	0.6260	0.6030
100	0.1795	0.6199	0.6347	0.6190
110	0.1423	0.6246	0.6367	0.6258
120	0.1149	0.6283	0.6367	0.6305
130	0.0934	0.6332	0.6400	0.6354
140	0.0775	0.6331	0.6360	0.6344
150	0.0653	0.6422	0.6427	0.6424
160	0.0566	0.6463	0.6440	0.6451
170	0.0498	0.6336	0.6300	0.6315
180	0.0449	0.6436	0.6387	0.6407
190	0.0404	0.6299	0.6227	0.6254
200	0.0370	0.6440	0.6353	0.6384

In accordance with the testing results, it seemed that an ordinary CNN did not fit enough for this dataset even though each trained epoch showed its convergence, but the validation results were not good. This was caused either by the difference between distribution of words in the training data and the one in testing data or an ordinary CNN model found difficulty to distinguish two sentences with very similar structures, but in fact they were different semantically. For instance, "What's your favorite political bumper sticker?" with "What was your favorite bumper sticker in the 2000s?". These two sentences was taged "not similar", even their structures were similar, and there were some others. Conceptually, a CNN usually performs convolution with kernel. The kernel size used in this model was [1, 3, 5] multiplied by the size of embedding yielded by Word2Vec. If there are similar sentence structures, the result of the convolution will only be different in some elements of convolution's results matrix. So that, the similarity between matrices of convolution's results from two sentence inputs will be high (i.e., two sentences are similar). Even though dropout and regularization had been implemented, this problem would still happened, because there were pair of sentences with similar structures would be considered either similar or not similar. In order to solve the problem it needs a language model and a more complex network. In this experiment, Word2Vec yielded vector of word based on surrounding words. It was not contextual, it means that one word would be represented by a vector that always the same.

From both Table I and Table II was derived chart comparing the two loss functions (triplet and contrastive) each for metric *precision*, *recall* and F1-*score*. Fig. 11, 12 and 13 show the chart of comparing metric between two loss functions, respectively for *precision*, *recall* and F1-*score*.

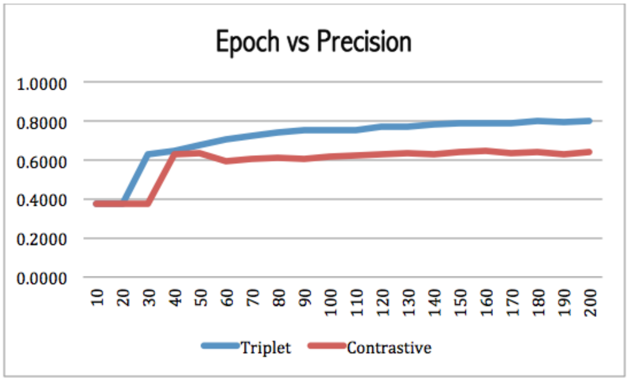


Fig. 11. The Chart of *Epoch* versus *Precision* for Triplet and Contrastive Loss Function

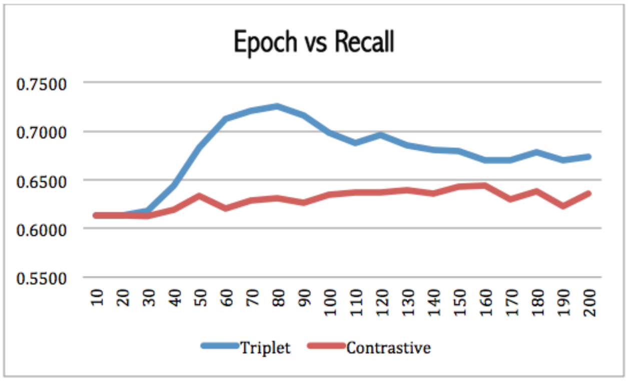


Fig. 12. The Chart of *Epoch* versus *Recall* for Triplet and Contrastive Loss Function

The three metric, *precision*, *recall* and F1-*score* were

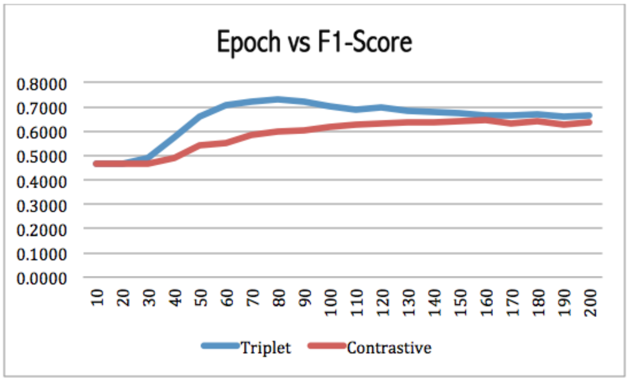


Fig. 13. The Chart of *Epoch* versus F1-*Score* for Triplet and Contrastive Loss Function

computed using the formula derived from a confusion matrix.

V. Conclusions

The two CNN **Siamese** models had been succesfully built, one with **triplet** loss function and the other with **contrastive** loss functions. The size of dataset used for training and testing were 3500 and 1500 respectively for the total of 5000. Both models were treated equally either in training or testing processes. In the training process, the model with **triplet** loss function consisted of three exact the same CNN models, while one with **contrastive** loss function only consisted of two exact the same CNN models. From the results of the model testing, it was seen that among 20 values of epoch there was one with epoch value of 180 with *precision* 0.8004, *recall* 0.6780, and F1-*score* 0.6713 for **triplet** loss function; and epoch value of 160 with *precision* 0.6463, *recall* 0.6440, and F1-*score* 0.6451 for **contrastive** loss function gave the best performance. So that, the **triplet** loss function gave better influence than **contrastive** loss function in measuring similarity between two given sentences.

References

[1] Bird, S., at.al, 2009, Natural Language Processing with Python, Published by O'Reilly Media, Inc., 105 Gravenstein Highway North, Sebastopol, CA 95472.

[2] Bowman, S. R., Angeli, G., Potts, C. and Manning, C. D., 2015, "A large annotated corpus for learning natural language inference". In Proceeding of the 2015 on Empirical Methods in Natural Language Processing, pages 632 - 642, Lisbon, Portugal. Association for Computational Linguistics.

[3] Bravo-Marquez, F., Mendoza, M., and Poblete, B., Meta-level sentiment models for big social data analysis, Knowl.-Based Syst. 69 (2014) 86–99, http://dx.doi. org/10.1016/j.knosys.2014.05.016.

[4] Chauhan, V. K., Bansal, A. and Goel, A., Twitter Sentiment Analysis Using Vader, International Journal of Advance Research, Ideas and Innovations in Technology. vol. 4, 1 (2018), 485-489.

[5] Chen, Q., Zhu, X., Ling, Z. H., Wei, S., Jiang, H. and Inkpen, D., 2017, "Enhanced LSTM for Natural Language Inference". In Proceedings of The 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages: 1657 - 1668; Vancouver, Canada. Association for Computational Linguistics.

[6] Gomma, W. H. and Fahmy, A. A., "A Survey of Text Similarity Approaches," Int. J. Comput. Appl., vol. 68, no. 13, 2013, doi: https://doi.org/10.5120/11638 - 7118.

[7] Gong, Y., Luo, H. and Zhang, J., 2018, "Natural Language Inference over Interaction Space", In Proceedings of the 6th International Conference on Learning Representations.

[8] Hidayat, E. Y., Firdausillah, F., Hastuti, K., Dewi, I. N. and Azhari, A., "Automatic Text Summarization Using Latent Drichlet Allocation (LDA) for Document Clustering," Int. J. Adv. Intell. Informatics, vol. 1, no. 3, p. 132, Dec. 2015, doi: https://doi.org/10.26555/ijain.v1i3.43.

[9] Hoffer, E., and Ailon, N., "Deep Metric Learning using Triplet Network" Accepted as a workshop contribution at ICLR 2015.

[10] Hutto, C. J. and Gilbert, E., VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Copyright © 2014, Association for the Advancement of Artificial Intelli-gence (www.aaai.org).

[11] Khot, T., Sabharwal and Clark, P., 2018. SciTail: A Textual Entailment Dataset from Science Question Answering. In Proceedings of AAAI.

[12] Kim, S., Hong, J. H., Kang, I. and Kwak, N., 2018, "Semantic Sentence Matching with Densely-connected Recurrent and Co-attentive Information". Computing Research Repository, arXiv: 1805.11360. Version 2.

[13] Liu, B.and Zhang, L., A survey of opinion mining and sentiment analysis, in: Mining Text Data, Springer, 2012, pp. 415–463, http://dx.doi.org/10.1007/978-1- 4614-3223-4_13.

[14] Liu, X., Duh, K. and Gao, J., 2018, "Stocastic Answer Networks for Natural Language Inference". Computing Reseach Repository, arXiv: 1804.07888.

[15] Parikh, A., T., Tackstrom, O., Das, D. and Uszkoreit, J., 2016, "A Decomposable Attention Model for Natural Language Inference". In Proceedings of the 2016 Conference of Empirical Methods in Natural Language Processing (EMNLP), pages 1532 - 1543, Doha, Qatar. Association for Computational Linguistics.

[16] Ravi, K., and Ravi, V., A survey on opinion mining and sentiment analysis: Tasks, approaches and applications, Knowl.-Based Syst. 89 (2015) 14–46, http://dx. doi.org/10.1016/j.knosys.2015.06.015.

[17] Wang, S. and Jiang, J., 2017. "A Compare-Aggregate Model for Matching Text Sequences". In Proceedings of the 5th International Conference on Learning Representations.

[18] Yang, R., Zhang, J., Gao, X., Ji, F. and Chen, H., 2019, "Simple and Efective Text Matching with Richer Alignment Features". In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4699 - 4709, Florence, Italy, July 28 - August 2, 2019. Association for Computational Linguistics.

[19] Yang, Y., Yih, W. and Meek, C., 2015, "WikiQA: A Chalenge Dataset for Open-domain Question Answering." In Proceeding of The 2015 Conference on Empirical Methods in Natural Language Processing", pages 2013 - 2018, Lisbon, Portugal. Association for Computational Linguistics.

[20] Yu, L., Hermann, K. M., Blunsom, P. and Pulman, S., 2014, "Deep Learning for Answer Sentence Selection". In NIPS Deep Learning and Representation Learning Work shop, Montreal.

[21] Yunianta, A., Barukah, O.M., Yusof, N., Dengen, N., Haviluddin, H. and Othman, S., "Semantic data mapping technology to solve semantic data problem on heterogeneity aspect," Int. J. Adv. Intell. Informatics, vol. 3, no. 3, pp. 161 - 172, Dec. 2017, doi: https://doi.org/10.26555/ijain.v3i3.131.

[22] https://miro.medium.com/max/2800/0*0efgxnFIaLTZ2qkY, "The Architecture of CNN".

[23] https://www.researchgate.net/publication/337578683_A_Survey_of_Vehicle_Re-Identification_Based_on_Deep_Learning/figures?lo=1, "Schematic diagram of the Siamese Network structure".

Comparative Evaluation of CNN Architectures for Image Caption Generation

Sulabh Katiyar[1], Samir Kumar Borgohain[2]
Department of Computer Science and Engineering
National Institute of Technology, Silchar
Assam, India 788010

***Abstract*—Aided by recent advances in Deep Learning, Image Caption Generation has seen tremendous progress over the last few years. Most methods use transfer learning to extract visual information, in the form of image features, with the help of pre-trained Convolutional Neural Network models followed by transformation of the visual information using a Caption Generator module to generate the output sentences. Different methods have used different Convolutional Neural Network Architectures and, to the best of our knowledge, there is no systematic study which compares the relative efficacy of different Convolutional Neural Network architectures for extracting the visual information. In this work, we have evaluated 17 different Convolutional Neural Networks on two popular Image Caption Generation frameworks: the first based on Neural Image Caption (NIC) generation model and the second based on Soft-Attention framework. We observe that model complexity of Convolutional Neural Network, as measured by number of parameters, and the accuracy of the model on Object Recognition task does not necessarily co-relate with its efficacy on feature extraction for Image Caption Generation task. We release the code at https://github.com/iamsulabh/cnn_variants.**

Keywords*—Convolutional Neural Network (CNN); image caption generation; feature extraction; comparison of different CNNs

I. Introduction

Image Caption Generation involves training a Machine Learning model to learn to automatically produce a single sentence description for an image. For human beings it is a trivial task. However for a Machine Learning method to be able to perform this task, it has to learn to extract all the relevant information contained in the image and then to convert this visual information into a suitable representation of the image which can be used to generate a natural language sentence description of the image. The visual features extracted from the image should contain information about all the relevant objects present in the image, the relationships among the objects and the activity settings of the scene. Then the information needs to be suitably encoded, generally in a vectorized form, so that the sentence generator module can convert this into a human readable sentence. Furthermore, some information may be implicit in the scene such as a scene where a group of football players are running in a football field but the football is not present in the scene frame. Thus the model may need to learn some level of knowledge about the world as well. However, the ability to automate the caption generation process has many benefits for the society as it can either replace or complement any method that seeks to extract some information from the images and has applications in the fields of education, military, medicine, etc. as well as applications in some specific problems such as helping visually impaired people in navigation or generating news information from images.

During the last few years there has been tremendous progress in Image caption generation due to advances in Computer Vision and Natural Language Processing domains. The progress made in Object Recognition task due to availability of large annotated datasets such as ImageNet [1] has led to availability of pre-trained Convolutional Neural Network (CNN) models which can extract useful information from the image in vectorized form which can then be used by caption generation module (called the decoder) to generate caption sentences. Similarly, progress in solving machine translation with methods such as encoder-decoder framework proposed in [2], [3] has led to adoption of similar format for Image Caption Generation where the source sentence in machine translation task is replaced by the image in caption generation task and then the process is approached as 'translation' of image to sentence, as has been done in works such as [4], [5], [6]. The attention based framework proposed by [7] where the decoder learns to focus on certain parts of the source sentence at certain time-steps has been adapted in caption generation in such as way that the decoder focuses on portions of image at certain time-steps [8]. A detailed survey of Image Caption Generation has been provided in [9] and [10].

Although there has been a lot of focus on the decoder which 'interprets' the image features and 'translates' them into a caption, there has not been enough focus on the encoder which 'encodes' the source image into a suitable visual representation (called image features). This is mainly because most methods use transfer learning to extract image features from pre-trained Convolutional Neural Networks (CNN) [11] which are trained on the Object Detection task of the ImageNet Large Scale Visual Recognition Challenge [12] where the goal is to predict the object category out of 1000 categories annotated in the dataset. Since the last layer of the CNN produces a 1000 length vector containing relative probabilities of all object categories, the last layer is dropped and the output(s) of intermediate layer(s) is(are) used as image features. Numerous CNN architectures have been proposed with varying complexity and efficacy and many have been utilized for Image Caption Generation as well. However, this makes it difficult to undertake a fair comparison of Image Caption Generation methods since the difference in performance could be either due to difference in effectiveness of decoders in sentence generation or due to difference in effectiveness of encoders in feature extraction.

Hence, in this work we evaluate Image Caption Generation using popular CNN architectures which have been used for Object Recognition task and analyse the co-relation between model complexity, as measured by the total number of parameters, and the effectiveness of different CNN architectures on feature extraction for Image Caption Generation. We use two popular Image Caption Generation frameworks: (a) Neural Image Caption (NIC) Generator proposed in [6] and (b) Soft Attention based Image Caption Generation proposed in [8]. We observe that the performance of Image Caption Generation varies with the use of different CNN architectures and is not directly correlated with either the model complexity or performance of CNN on object recognition task. To further validate our findings, we evaluate multiple versions of ResNet CNN [13] with different depths (number of layers in the CNN) and complexity: ResNet18, ResNet34, ResNet50, ResNet101, ResNet152 where the numerical part in the name stands for the number of layers in the CNN (such as 18 layers in ResNet18 and so on). We evaluate multiple versions of VGG CNN [14] architecture: VGG-11, VGG-13, VGG-16 and VGG-19 and multiple versions of DenseNet CNN [15] architecture: Densenet121, Densenet169, DenseNet201 and Densenet161, each of which has different number of parameters. We observe that performance does not improve with the increase in the number of layers, and consequently, increase in model complexity. This further validates our observation that effectiveness of CNN architectures for Image Caption Generation depends on the model design and that the model complexity or the performance on Object Detection task are not good indicators of effectiveness of CNN for Image Caption Generation. To the best of our knowledge, this is the first such detailed analysis of the role of CNN architectures as image feature extractors for Image Caption Generation task. In addition, to further the future research work in this area, we also make the implementation code[1] available for reference.

This paper is divided into following sections: In Section II,we discuss the relevant methods proposed in the literature, in Section III, we discuss the methodology of our work, in Section IV we present and discuss the experimental results and in Section V we discuss the implications of our work and possible future studies.

II. Related Work

Some of the earliest works attempted to solve the problem of caption generation in constrained environments such as the work proposed in [16] where the authors try to generate captions for objects present in an office setting. Such methods had limited scalability and applications. Some works tried to address the task as a *Retrieval problem* where a pool of sentences was constructed which could describe all (or most) images in a particular setting. Then for a target image, a sentence which was deemed appropriate by the algorithm was selected as the caption. For example, in [17], the authors construct a 'meaning space' which consists of triplets of <objects, actions, scene>. This is used as a common mapping space for images and sentences. A similarity measure is used to find sentences with the highest similarity to the target image and the most similar sentence is selected as the caption. In [18], a set of images are retrieved from the training data which are similar to the target image using a visual similarity measure. Then a word probability density conditioned on the target image is calculated using the captions of the images that were retrieved in the last step. Then the captions in the dataset are scored using this word probability density and the sentence which has the highest score is selected as the caption for the target image. The retrieval based methods generally produce grammatically correct and fluent captions because they select human generated sentence for a target image. However, this approach is not scalable because a large number of sentences need to be included in the pool for each kind of environment. Also the selected sentence may not even be relevant because the same kind of objects may have different kind of relationships among them which cannot be described by a fixed set of sentences.

Another class of approaches are the *Template based methods* which construct a set of hand-coded sentence templates according to the rules of grammar and semantics and optimization algorithms. Then the methods plug in different object components and their relationships into the templates to generate sentences for the target image. For example, in [19], Conditional Random Fields are used to recognize image contents. A graph is constructed with the image objects, their relationships and attributes as nodes of the graph. The reference captions available with the training images are used to calculate pairwise relationship functions using statistical inference and the visual concepts are used to determine the unary operators on the nodes. In [20], visual models are used to extract information about objects, attributes and spatial relationships. The visual information is encoded in the form of [<adjective1,object1>,preposition,<adjective2,object2>] triplets. Then n-gram frequency counts are extracted from web-scale training dataset using statistical inference. Dynamic programming is used to determine optimal combination of phrases to perform phrase fusion to construct the sentences. Although the Template based approaches are able to generate more varied captions, they are still handicapped by the problems of scalability because a large number of sentence templates are to be hand-coded and even then a lot of phrase combinations may be left out.

In recent years, most of the works proposed in the literature have employed Deep Learning to generate captions. Most works use CNNs, which are pre-trained on the ImageNet Object Recognition dataset [1], to extract vectorized representation of the image. Words of a sentence are represented as Word Embedding vectors extracted from a look-up table. The look up table is learned during training as the set of weights of the Embedding Layer. The image and word information is combined in different ways. Most methods use different variants of Recurrent Neural Network [21] (RNN) to model the temporal relationships between words in the sentence. In [5], the image features extracted from CNN and the word embeddings are mapped to the same vector space and merged using element-wise addition at each time-step. Then the merged image features and word embeddings are used as input to a MultiModal Reccurent Neural Network (m-RNN) which generates the output. The authors use AlexNet[22] and VGG-16 [14] as CNNs to extract image features. In [4] a Bidirectional Recurrent Neural Network is used as decoder because it can map the word relationships with both the words that precede and the words that succeed a particular word in

[1] https://github.com/iamsulabh/cnn_variants

the sentence. The word embeddings and image features are merged before being fed into the decoder. The authors use AlexNet [22] CNN to extract image features. In [6], a Long Short Term Memory Network [23] is used as decoder. The image features are mapped to the vector space spanned by hidden state representations of the LSTM and are used as initial hidden state of the LSTM. Thus the image information is fed to LSTM at initial state only. The LSTM takes in previously generated words as input (with a special 'start token' as the first input) and generates the next word sequentially. The authors use [24] as CNN for extracting image features. Using the Attention approach, in [8] the authors train the model to focus on certain parts of the image at certain time-steps. This attention mechanism takes as input, the image features and output until the last time-step and generates an image representation conditioned on text input. This is merged with the word embeddings at the current time-step by using vector concatenation operation and used as input to the LSTM generator. The authors used VGGNet [14] CNN as image feature extractor. Recently, methods using Convolutional Neural Networks as sequence generators have been proposed such as in [25] for text generation. Based on this approach, [26] propose a method which uses a CNN for encoding the image and another CNN for decoding the image. The CNN decoder is similar to the one used in [25] and uses a hierarchy of layers to model word relationships of increasing complexity. The authors use ResNet152[13] CNN to encode the image features. More recently, Transformer Network has been used which uses self-attention to model word relationships instead of Recurrent or Convolutional operations [27]. Based on this approach a Transformer based caption generation is proposed in [28]. Since most of the methods use different CNN architectures to extract image features, there is a need for a comparative analysis of their effectiveness in image feature extraction using the same overall format for caption generation.

III. Proposed Method

In image caption generation, given an image the task is to generate a set of words $S = \{w_1, w_2, w_3, ..., w_L\}$ where $w_i \in \mathcal{V}$ where L is the length of the sentence and $\mathcal{V}$ represents the vocabulary of the dataset. The words w_1 and w_L are usually the special tokens for start and end of the sentence. Two more special tokens for 'unknown' and 'padding' are also used for representing unknown words (which may be the stop words and rare words that have been removed from dataset to speed up training) and padding the end of the sentence (to make all sentences of equal length because RNNs do not handle sentence of different lengths in the same batch), respectively. Given pairs of image and sentence, (I_N, S_i) for $i \in (1, 2, 3, ..., j)$, during training we maximize the probability $P(S_i|I_N, \theta)$ where j is the number of captions for an image in training set and θ represents the set of parameters in the model. Hence, as mentioned in [6], during training the model learns to update the set of parameters θ such that the probability of generation of correct captions is maximized according to the equation,

$$\theta^\star = argmax \sum_{(I,S)} log_p(S|I, \theta) \tag{1}$$

where θ is the set of all parameters of the model, I is the image and S is one of the reference captions provided with the image. We can use chain rule because generation of words of a sentence depends on previously generated words, and hence Equation 1 can be extended to the constituent words of the sentence as,

$$log_p(S|I, \theta) = \sum_{t=0}^{L} log_p(w_t|I, \theta, w_1, w_2, ..., w_{t-1}) \tag{2}$$

where $w_1, w_2, ..., w_L$ are the words in the sentence 'S' of length L. This equation can be modelled using a Recurrent Neural Network which generates the next output conditioned on the previous words of the sentence. We have used LSTM as the RNN variant for our experiments.

In this work, we evaluate caption generation performance on two popular encoder-decoder frameworks with certain modifications. For both the methods, we experiment with different CNN architectures for image feature extraction and analyse the effects on performance.

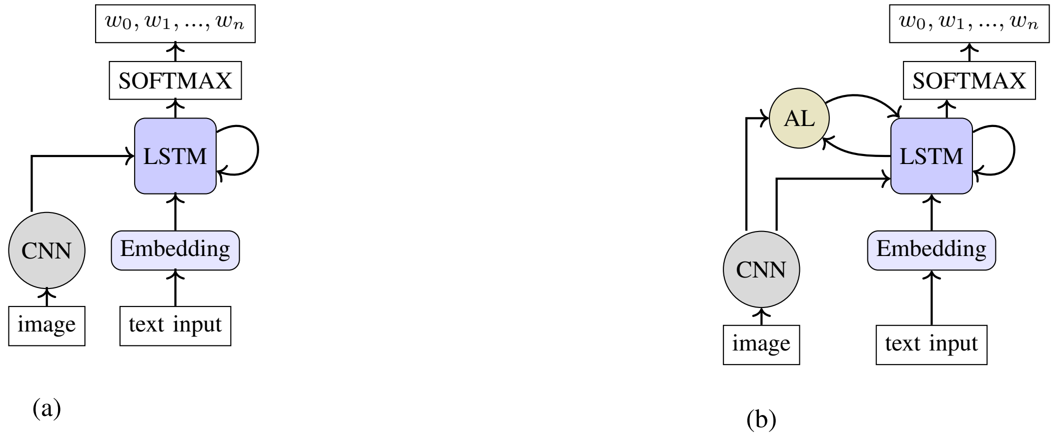
The first method is based on Neural Image Caption Generation method proposed in [6]. However, unlike the method proposed in [6], we have not used model ensembles to improve performance. In addition, we have extracted image features from a lower layer of the CNN which generates a set of vectors each of which contain information about a region of the image. We have observed that this leads to better performance as the decoder is able to use region specific information to generate captions. Throughout this paper, this will be referred to as 'CNN+LSTM' approach with the word 'CNN' replaced by the name of CNN architecture used in the experiment. For example, 'ResNet18+LSTM' refers to caption generation with ResNet18 as the CNN.

The second method is similar to the Soft Attention method proposed in [8]. We use an attention mechanism which learns to focus on certain portions of image for at certain time-steps for generating the captions. Similar to the CNN+LSTM approach, this Soft Attention approach will be referred as 'CNN+LSTM+Attention' approach with the word 'CNN' replaced by the name of CNN architecture used. Figure 1 explains both the methods.

A. Image Feature Extraction

For extracting image features, we use CNNs which were pre-trained on ImageNet datset [1] for the Imagenet Large Scale Visual Recognition Challenge [12]. The models generate a single output vector containing the relative probabilities of different object categories (with 1000 categories in total). We remove this last layer from the CNN since we need more fine-grained information. Also, we remove all the layers at the top (with the input layer being called the bottom layer) which produce a single vector as output because we need a set of vectors as output which contain information about different regions of the image. Hence, the image features are a set of vectors denoted as, $\mathbf{a} = \{a_1, a_2, a_3, ...a_{|a|}\}, a_i \in \mathcal{R}^D$ where $|a|$ is the number of feature vectors contained in $\mathbf{a}$, $\mathcal{R}$ represents real numbers and D is dimension of each vector. For example, ResNet152 CNN [13] generates a set of 8, 2048 dimensional vectors.

The set of image feature vectors thus generated are used in two ways in the methods used in this work. In the 'CNN+LSTM' method, the image features are mapped to



(a)

(b)

Fig. 1. An Overview of the Two Approaches Proposed in this Work: (a) Encoder-Decoder based Approach. (b) Attention based Approach with an Attention Mechanism to Focus on Salient Portions of the Image. (AL stands for Attention Layer)

the vector space of hidden state of the LSTM and used to initialize the hidden and cell state of the LSTM decoder. For the 'CNN+LSTM+Attention' method, in addition to hidden and cell state initialization, the set of image feature vectors is also used at each time-step to calculate attention weighted image features which contain information from those regions in the image which are important at the current time-step. We explain this in detail in Sections III-B and III-C.

B. CNN + LSTM Method

In this method, we use a CNN encoder to extract image information and use that information as the initial hidden state of the LSTM decoder. Using the set of image feature vectors thus obtained as described in Section III-A, we obtain a single vector by averaging the values of all vectors in the set as,

$$a_{ave} = \sum_{i}^{|\mathbf{a}|} a_i, i \in (1, 2, ..., |\mathbf{a}|) \tag{3}$$

where $|\mathbf{a}|$ is the length of set of image feature vectors extracted from the CNN. This is used to generate the initial hidden and cell states of the LSTM by using an affine transformation followed by a non-linearity ($Tanh$ function) as,

$$h_0 = Tanh(a_{ave} \star W^h + b^h) \tag{4}$$

$$c_0 = Tanh(a_{ave} \star W^c + b^c) \tag{5}$$

where W^h, W^c and b^h, b^c are weights and biases of the MultiLayer Perceptron (MLP) which is used to model the transformations.

The successive hidden and cell states are generated during training. Since the generation of words is dependent on the previous words in the sentence as depicted in Equation 2, this dependence can be modelled using the hidden state of the LSTM (which is also modulated by the cell state). Hence,

$$P_\theta(w_i|I, w_1, w_2, ..., w_{i-1}) = P_\theta(w_i|I, h_i) = f_\theta(w_i, I, h_i) \tag{6}$$

where f_θ is any differentiable function and since it is recursive in nature it can be modelled using an RNN. Since the hidden state also depends on the previous hidden states, it can be modelled as a function of previous hidden state and inputs as,

$$h_i = f_\theta(w_{i-1}, h_{i-1}, I) \tag{7}$$

where f_θ is the same differentiable function as in Equation 6 since the model is trained end-to-end with the same parameters. And words are represented as word embeddings which is a function that maps one-hot word vectors to the embedding dimensions and is also learned with the rest of the model, as

$$w_i^e = f_\theta(w_i) \tag{8}$$

where f_θ is the same differentiable function in Equation 6 and w_i^e is the word embedding vector for word w_i.

We use LSTM as described in [23]. The LSTM has three control gates: input, forget and update gates. The equations for updating the different gates are as follows:

$$i_t = \sigma(W_i x_t + R_i h_{t-1} + b_i) \tag{9}$$

$$f_t = \sigma(W_f x_t + R_f h_{t-1} + b_f) \tag{10}$$

$$o_t = \sigma(W_o x_t + R_o h_{t-1} + b_o) \tag{11}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_z x_t + R_z h_{t-1} + b_z) \tag{12}$$

$$h_t = o_t \odot tanh(c_t) \tag{13}$$

where W_i and R_i, W_f and R_f, W_o and R_o and W_z and R_z are weight matrices (input and recurrent weight matrices) pairs for the input, forget, output and the input modulator(tanh) gates, respectively. b is the bias vector and σ is the sigmoid function. It is expressed as $\sigma(x) = 1/1{+}exp(x)$ and condenses the input to the range of (0,1). $\tanh$ is the is hyperbolic tangent function which condenses the input in the range (-1,1). i_t, o_t and f_t are input, output and forget gates respectively. The input gate processes the input information. The output gate generates output based on the input and some of this information has to be dropped which is decided by the cell state. The cell state stores information about the context. The forget gate decides what contextual information has to be dropped from the cell state. The internal structure of the LSTM has been depicted in Fig. 2.

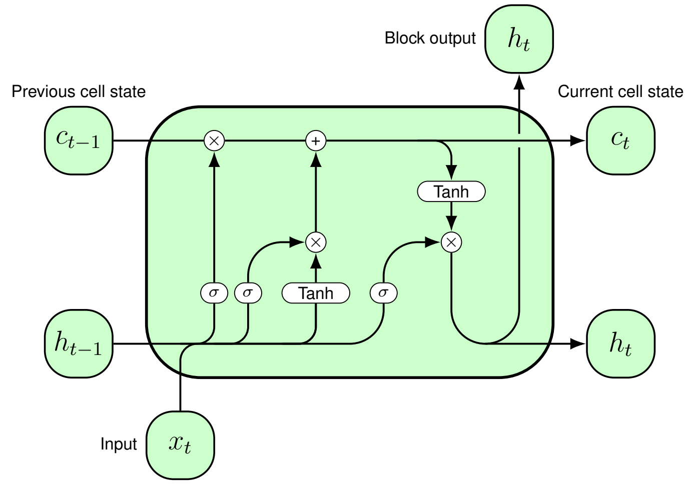


Fig. 2. Illustration of a basic LSTM Cell.

C. CNN + LSTM + Attention Method

In this method, in addition to the the initial time-step, the image information is fed into the LSTM at each time-step. However a separate attention mechanism generates information which is extracted from only certain regions of image which are relevant at the current time-step.

The attention mechanism produces a context vector which represents the relevant portion of the image at each time-step. First a set of weights are calculated for each image feature vector $a_i \in \mathbf{a}, i \in (1, 2, 3, ..., |a|)$ as described in Section III-A.

$$P = \{p_{ti}\}, \qquad p_{ti} = f_{att}(a_i, h_{t-1}) \tag{14}$$

where $i \in (1, 2, 3, ..., |a|)$. Then the attention weights are calculated as,

$$\boldsymbol{\alpha} = \{\alpha_{ti}\}, \qquad \alpha_{ti} = \frac{exp(p_{ti})}{\sum_{k=1}^{n} exp(p_{tk})} \tag{15}$$

where $\boldsymbol{\alpha}$ is the set of weights, one for each image feature vector a_i in $\mathbf{a}$ such that $\sum_{k=1}^{|a|} \alpha_i = 1$.

Then the context vector is calculated by another function,

$$\mathbf{z_i} = \Phi(\{a_i\}, \{\alpha_i\}) \tag{16}$$

We have used the function f_{att} and Φ as desrcibed in [8].

With the context vector thus obtained, the equations for the gates of the LSTM decoder would be,

$$i_t = \sigma(W_i x_t + R_i h_{t-1} + Z_i z_t + b_i) \tag{17}$$

$$f_t = \sigma(W_f x_t + R_f h_{t-1} + Z_f z_t + b_f) \tag{18}$$

$$o_t = \sigma(W_o x_t + R_o h_{t-1} + Z_o z_t + b_o) \tag{19}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_c x_t + R_c h_{t-1} + Z_c z_t + b_c) \tag{20}$$

$$h_t = o_t \odot tanh(c_t) \tag{21}$$

where W_i and R_i, W_f and R_f, W_o and R_o and W_c and R_c are weight matrices (input and recurrent weight matrices) pairs for the input, forget, output and the input modulator(tanh) gates, respectively. b is the bias vector and σ is the sigmoid function.

IV. Experiments and Results

In this section we describe the experimental details and the results. We have evaluated Squeezenet [31], Shufflenet [32], Mobilenet [33], MnasNet [34], ResNet [13], GoogLeNet [29], DenseNet [15], Inceptionv4 [24], AlexNet [22], DPN (Dual Path Network) [36], ResNext [37], SeNet [39], PolyNet [40], WideResNet [38], VGG [14], NASNet-Large [35] and InceptionResNetv2 [41] CNN models. Out of these we have evaluated five versions of ResNet, viz. Resnet18, ResNet34, ResNet50, Resnet101, Resnet152, four versions of DenseNet, viz. Densenet121, Densenet169, DenseNet201 and Densenet161 and four versions of VGG, viz. VGG-11, VGG-13, VGG-16 and VGG-19 which are similar in architecture but differ widely in terms of number of parameters and also in terms of accuracy and error rates on Object Recognition task with ImageNet dataset.

We have evaluated the performance using BLEU, METEOR, CIDER, ROUGE-L and SPICE metrics that were recommended in MSCOCO Image caption Evaluation task [42]. The evaluation results are provided in Tables I and II for ‘CNN+LSTM’ and ‘CNN+LSTM+Attention’ methods, respectively. In addition we have provided some examples of generated captions in Tables III and IV for both the methods.

We have used Flickr8k [30] dataset which contains around 8000 images with five reference captions each. Out of the 8000 images, around 1000 are earmarked for validation set, around 1000 are meant for test set and the remaining are for training set.

We can make following observations from the results:

- For example, there is a variation of around 4 to 5 points in the evaluation metrics between the best and worst performing models in both Tables I and II.
- In addition, the performance of a decoder framework which employs additional methods of guidance (such as attention) but uses a lower performing encoder can be worse than simpler methods which use better performing CNN encoder. For example, the best performing model using CNN+LSTM method (Table I) have better performance than lower performing models using CNN+LSTM+Attention method (Table II).
- Although different variants of the same model (such as ResNet, Densenet and VGG) differ greatly with respect to the number of parameters, they generate image captioning performances which differ only by around 1 point on most evaluation metrics. ResNet18, being the smallest model in terms of number of parameters (among ResNet based CNNs) performs competitively as compared to the larger ResNet variants which have many times more parameters. We also observe that DenseNet121 and VGG-11 being the smallest models among DenseNet and VGG models, respectively, outperform other DenseNet and VGG based CNNs in evaluation scores along certain metrics.
- Also the different variants of ResNet [13], VGG [14] and DenseNet [15] architectures differ greatly in terms

TABLE I. Performance of CNN+LSTM method using different CNN architectures.

CNN name	Parameters (in thousands)	Top-5 O.D. error	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDER	ROUGE-L	SPICE
Squeezenet [31]	1,248	19.58	60.04	40.65	26.95	17.61	18.12	42.87	44.05	12.44
Shufflenet[32]	2,279	11.68	59.70	41.18	27.84	18.67	18.24	44.36	43.66	12.61
Mobilenet[33]	3,505	9.71	60.60	41.72	28.44	18.87	18.83	47.97	44.28	13.50
MnasNet[34]	4,383	8.456	61.19	43.02	29.43	20.10	18.94	48.19	44.88	13.46
Densenet121 [15]	7,979	7.83	61.62	43.36	29.47	19.88	19.39	48.99	45.32	13.64
ResNet18 [13]	11,689	10.92	62.21	43.45	29.84	20.30	18.91	48.31	45.33	13.49
GoogLeNet [29]	13,005	10.47	60.69	41.57	28.20	18.91	18.66	46.42	44.38	13.01
Densenet169 [15]	14,150	7.00	63.73	45.00	30.87	21.13	19.95	52.88	46.41	14.32
DenseNet201 [15]	19,447	6.43	63.29	45.11	31.36	21.63	19.80	52.21	46.40	14.16
Resnet34 [13]	21,798	8.58	61.08	42.69	29.32	19.98	18.98	49.78	45.01	13.32
Resnet50 [13]	25,557	7.13	61.86	43.79	30.10	20.27	19.11	50.86	45.76	13.89
Densenet161 [15]	28,681	6.20	63.12	44.68	30.76	20.79	20.00	54.24	46.19	14.26
Inceptionv4 [24]	42,680	4.80	59.49	40.47	27.00	18.03	18.22	43.17	43.61	12.23
Resnet101 [13]	44,549	6.44	62.77	44.11	30.62	21.10	19.65	53.00	45.91	14.04
InceptionResNetv2 [41]	54,340	4.9	59.50	40.55	27.36	18.21	18.79	46.35	43.54	12.90
ResNet152 [13]	60,193	5.94	62.30	44.24	30.84	21.21	19.50	55.10	46.14	14.20
AlexNet [22]	61,101	20.91	59.24	40.17	26.82	17.87	17.51	41.09	42.79	11.78
DPN131 [36]	75,360	5.29	59.60	40.69	27.58	18.86	18.00	42.36	43.15	12.67
ResNext101 [37]	88,791	5.47	62.38	43.79	29.85	20.20	19.54	51.37	45.54	14.05
NASNetLarge [35]	88,950	3.8	56.08	36.76	23.54	15.46	16.76	34.74	40.50	11.56
SeNet154 [39]	115,089	4.47	61.67	43.18	29.72	20.19	19.48	49.89	45.24	13.95
PolyNet [40]	118,733	4.25	60.26	41.26	27.68	18.68	18.02	44.23	43.61	12.37
WideResNet101 [38]	126,886	5.72	61.42	42.48	28.71	19.16	18.64	46.24	44.41	13.23
VGG-11(bn) [14]	132,869	11.37	61.70	43.37	30.08	20.86	19.38	48.98	45.80	13.62
VGG-13(bn) [14]	133,054	10.75	60.79	42.42	28.91	19.70	19.06	46.57	44.84	13.39
VGG-16(bn) [14]	138,366	8.50	60.56	41.98	28.66	19.51	19.04	48.41	44.82	13.71
VGG-19(bn) [14]	143,678	9.12	61.40	43.09	29.49	20.02	19.15	49.42	45.43	13.61

TABLE II. Performance of CNN+LSTM+Attention method using different CNN architectures.

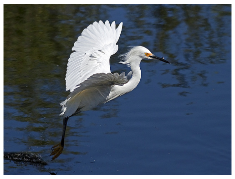
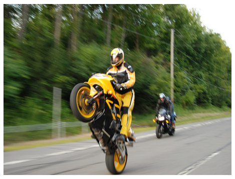
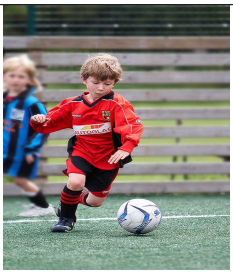
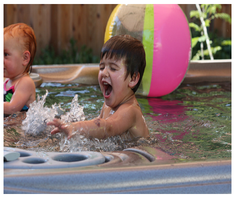
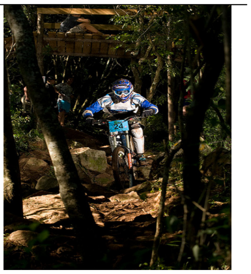
CNN name	Parameters (in thousands)	Top-5 O.D. error	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDER	ROUGE-L	SPICE
Squeezenet [31]	1,248	19.58	60.79	42.29	28.78	19.41	18.80	46.54	44.48	12.85
Shufflenet[32]	2,279	11.68	62.36	43.87	30.42	21.00	19.18	49.01	45.00	13.50
Mobilenet[33]	3,505	9.71	63.69	45.33	31.72	21.89	19.63	55.36	46.28	14.25
MnasNet[34]	4,383	8.456	63.99	45.75	32.11	22.36	19.78	54.84	46.17	14.02
Densenet121 [15]	7,979	7.83	64.11	45.67	31.76	22.07	20.43	55.85	46.74	14.91
ResNet18 [13]	11,689	10.92	63.26	44.87	31.07	21.24	20.08	52.44	45.84	13.75
GoogLeNet [29]	13,005	10.47	62.91	44.27	30.27	20.50	19.51	50.72	46.02	13.80
Densenet169 [15]	14,150	7.00	64.48	46.17	32.28	22.30	20.81	56.25	46.82	14.93
DenseNet201 [15]	19,447	6.43	64.38	46.26	32.41	22.49	20.73	59.71	47.19	15.13
Resnet34 [13]	21,798	8.58	63.36	45.28	31.88	22.23	19.88	55.35	46.17	14.40
Resnet50 [13]	25,557	7.13	65.32	46.92	32.81	22.58	20.87	57.12	46.95	14.90
Densenet161 [15]	28,681	6.20	65.00	46.99	32.83	22.56	20.44	56.74	47.57	14.93
Inceptionv4 [24]	42,680	4.80	60.17	42.24	28.71	19.35	18.76	48.00	44.33	13.26
Resnet101 [13]	44,549	6.44	64.33	45.99	32.13	22.02	20.29	56.09	46.58	14.80
InceptionResNetv2 [41]	54,340	4.9	61.46	42.98	29.20	19.84	19.20	49.83	44.44	13.81
ResNet152 [13]	60,193	5.94	65.26	47.55	33.72	23.67	20.94	58.33	47.54	15.18
AlexNet [22]	61,101	20.91	59.93	40.97	27.80	19.06	18.67	46.11	44.09	12.57
DPN131 [36]	75,360	5.29	62.68	44.17	30.47	20.53	19.41	49.98	45.51	13.95
ResNext101 [37]	88,791	5.47	64.78	46.07	32.36	24.45	20.93	57.67	40.04	15.28
NASNetLarge [35]	88,950	3.8	63.60	44.66	30.16	19.93	19.73	51.34	45.49	14.00
SeNet154 [39]	115,089	4.47	64.23	45.94	32.54	22.62	20.81	58.45	46.83	15.05
PolyNet [40]	118,733	4.25	62.56	44.78	31.16	21.48	19.75	53.38	45.96	13.81
WideResNet101 [38]	126,886	5.72	63.47	45.37	31.71	21.73	19.84	54.27	46.23	14.51
VGG-11(bn) [14]	132,869	11.37	63.00	44.66	31.18	21.68	19.79	52.24	46.42	14.08
VGG-13(bn) [14]	133,054	10.75	63.64	45.09	31.26	21.41	20.25	55.17	46.35	14.64
VGG-16(bn) [14]	138,366	8.50	63.81	45.77	32.35	22.55	20.19	55.13	46.72	14.49
VGG-19(bn) [14]	143,678	9.12	62.57	44.63	30.97	21.44	19.76	54.10	46.23	14.44

of Top-5 error on Object Detection task when evaluated with Imagenet dataset. However, that difference does not translate to similar difference in performance in Image captioning task.

- For each image, most models generate reasonable captions but there is a great variation in the caption sentences generated with different models. In some cases, captions generated with different models describe different portions of the image and sometimes some models focus on a certain object in the image instead of providing a general overview of the scene.
- In some cases, models do not recognize certain objects in the image. In particular, we have observed many cases of incorrect gender identification which points out to possible statistical bias in the dataset towards a particular gender in a certain context.

Thus we can conclude that choice of CNN for the encoder significantly influences the performance of the model. In addition to the general observations, we are able to deduce

TABLE III. EXAMPLES OF GENERATED CAPTIONS BY CNN+LSTM METHOD USING DIFFERENT CNN ARCHITECTURES.

Choice of CNN					
ResNet-152	a white crane flies over the water	a man riding a motorcycle	two young boys playing soccer	two children playing in a pool	a person riding a bike in the woods
Inception-ResNet	a white crane flies over the water	a man riding a motorcycle	two young boys playing soccer	two children are playing in a pool	a man in a blue shirt is riding a bike through a wooded area
NASNET Large	a white crane flies over the water	a man is riding a red motorcycle	a boy in a red uniform kicks a soccer ball	a child plays in a pool	a man on a bike in a forest
VGG-16	a white bird flies through the water	a man riding a yellow motorcycle	a boy in a soccer uniform kicking a soccer ball	a boy in a blue shirt plays with a plastic toy	a dirt bike rider is airborne in the woods
Alexnet	a white bird flies over the water	a man in yellow and yellow motorcycle	a boy in a red uniform runs with a soccer ball	a young girl in a bathing suit is jumping into a pool	a man is riding a bike on a dirt path
Squeezenet	a white bird in the water	a man in a yellow helmet is riding a bike	a boy in a red and white uniform is playing soccer	a little girl in a pink dress is playing in a pool	a person riding a bike through the woods
Densenet-201	a white bird flies over the water	two bikers racing on the road	two children playing soccer	a young boy in a pool	a man on a bike is riding a bike through the woods
GoogLeNet	a white bird flies through the water	a man on a motorcycle is riding on a street	a young boy wearing a red shirt and a blue soccer ball	a little boy is being splashed in a pool	a man is riding a bike through the woods
Shufflenet	a white bird flies through the water	a man in a yellow helmet riding a yellow bike	a little boy in a red shirt is playing with a soccer ball	two young children playing in a fountain	a man in a blue helmet rides a bike through the woods
Mobilenet	a white bird is flying over water	a person riding a bike in a race	a boy in a red and white uniform is playing soccer	a young boy in a swimming pool	a person riding a dirt bike in the woods
Resnext-101	a white bird flies over the water	a man on a motorcycle is riding a motorcycle	a soccer player in a red uniform kicks a soccer ball	a little girl is playing in a pool	a dirt bike rider in the woods
Wide ResNet-101	a white bird flies over the water	a man riding a motorcycle	two boys playing soccer on a field	a boy is splashing in a pool	a person riding a dirt bike through the woods
Mnasnet	a white bird in the water	a man in a yellow jacket rides a motorcycle	a boy in a blue uniform is playing soccer	a little boy is playing in a pool	a man on a bike in the woods
Inception	a white bird flying over water	a man is riding a bike on a track	two boys playing soccer	two children play in a pool	a person in a blue shirt and blue jeans is sitting on a tree
DPN-131	a white crane landing in the water	a person on a motorcycle	a young boy in a soccer uniform kicking a soccer ball	a little boy in a swimming pool	a person is riding a bike in the woods
Senet-154	a white crane flying over water	a man is riding a yellow motorcycle	a man in a red uniform kicking a soccer ball	a little boy in a swimming pool	a person rides a bike through the woods
Polynet	a white bird flies over the water	a man rides a motorcycle	a boy in a blue uniform is chasing a soccer ball	a girl in a pink shirt is playing in a kiddie pool	a person rides a bike through the woods

the following specific observations about the choice of CNN:

- ResNet [13] and DenseNet [15] CNN architectures are well suited to Image caption generation and generate better results while having a lower model complexity than other architectures.

V. CONCLUSION

In this work, we have evaluated encoder-decoder and attention based caption generation frameworks with different choices of CNN encoders and observed that there is a wide variation in terms of both the scores, as evaluated with commonly used metrics (BLEU, METEOR, CIDER, SPICE, ROUGE-L), and also the generated captions while using different CNN encoders. In terms of most metrics, there is a difference in performance of around 4-5 points between the worst and best performing models. Hence, the choice of particular CNN architecture plays a big role in the image caption generation process. In particular, ResNet and DenseNet based CNN architectures lead to better overall performance while at the same using lesser parameters than other models.

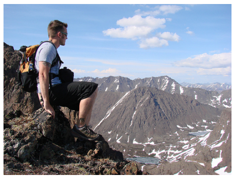
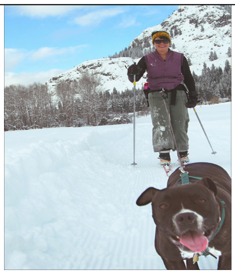
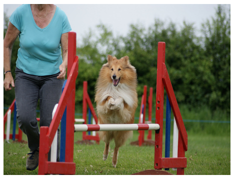
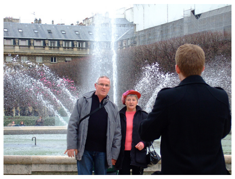
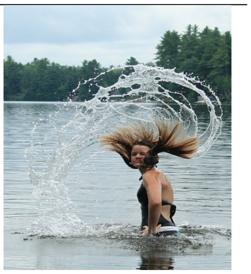
Also, since there is a great variation in the generated captions for each image, it may be possible to use ensemble of models, each of which utilize a different CNN as encoder, to increase diversity of generated captions. Also, model ensembling would lead to better performance. In the works proposed in the literature, model ensembling has been used such as in [6] but such model ensembles utilize similar models trained with different hyperparameters. Using ensembles of models, which use different CNN encoders is an area which could be explored in future works.

Furthermore, we hope that this analysis of the effect of choice of different CNNs for image captioning will aid the researchers in better selection of CNN architectures to be used as encoders in image feature extraction for Image Caption Generation.

ACKNOWLEDGMENT

We are greatly indebted to the MultiMedia Processing and Language Processing Laboratories at the Department of Computer Science and Engineering, National Institute of

TABLE IV. EXAMPLES OF GENERATED CAPTIONS BY CNN+LSTM+ATTENTION METHOD USING DIFFERENT CNN ARCHITECTURES.

Choice of CNN					
ResNet-152	a man is standing in front of a mountain	a dog runs through the snow	a dog jumps over a hurdle	a man and a woman are sitting on a fountain	a young girl in a pink bathing suit is playing in the water
Inception-ResNet	a man with a backpack stands on a mountaintop	a man in a red jacket is skiing down a snowy hill	a brown dog jumps over a hurdle	two children playing in a fountain	a little girl plays in the water
NASNET Large	a man sits on top of a mountain	a dog is running through the snow	a brown dog is jumping over a hurdle	a group of people are playing in a fountain	a girl in a pink swimsuit is jumping into the water
VGG-16	a man is standing on top of a mountaintop	a brown dog is standing in the snow	a dog is jumping over a hurdle	a group of people are sitting on a ledge overlooking a city	a woman in a swimsuit is standing in the water
Alexnet	a man is standing on top of a mountain	a brown dog is running through snow	a dog jumps over a hurdle	a man in a black jacket is standing next to a building	a boy in a pool
Squeezenet	a group of people sit on a snowy mountain	a man in a red jacket is standing on a snowy hill	a brown and white dog with a red and white dog	a group of people stand in front of a building	a woman in a white shirt is walking through the water
Densenet-201	a man in a blue shirt is standing in the mountains	a brown dog is jumping in the snow	a dog jumps over a hurdle	a man and a woman are standing in front of a fountain	a young girl jumping into the water
GoogLeNet	a man is standing on a mountaintop	a black and white dog is running through the snow	a dog jumps over a hurdle	a group of people stand in a fountain	a young boy plays in the water
Shufflenet	a man and a woman are sitting on a rock overlooking the mountains	a man in a red jacket is standing on a snowy hill	a woman and a dog are playing in a yard	a man and a woman are walking down a city street	a man is standing on the shore of a body of water
Mobilenet	a man stands on a mountain	a man is skiing down a snowy hill	a woman and a woman sitting on a bench	two men are standing next to a fountain	a girl in the water
Resnext-101	a man with a backpack stands on a mountaintop	a person is skiing down a snowy hill	a dog jumping over a hurdle	a man and a woman are standing in a fountain	a woman in a bikini is playing in the water
Wide ResNet-101	a man is standing on top of a mountain	a dog is running through the snow	a man and a dog on a leash	a group of people are standing in a fountain	a woman in a bathing suit walks along the water
Mnasnet	a man and a woman are standing in the mountains	a brown dog is running through the snow	a dog jumps over a hurdle	a group of people are standing in front of a fountain	a boy is splashing in the water
Inception	a man stands on a rock overlooking the mountains	a black and white dog in the snow	a brown and white dog is jumping over a hurdle	a group of people are playing in a fountain	a dog walks through the water
DPN-131	a man is standing on top of a mountain	a man and a dog play in the snow	a dog jumps over a hurdle	a group of people stand in a fountain	a girl in a swimsuit is jumping into the water
Senet-154	a man is standing in front of a mountain	a dog is running through the snow	a dog jumps over a hurdle	a man is standing in front of a fountain	a girl in a red bathing suit splashes in the water
Polynet	a man stands on a mountaintop	a dog is jumping over a snowy hill	a dog is jumping over a hurdle	a group of people are standing in front of a fountain fountain	a woman in a bathing suit is standing in front of a waterfall

Technology, Silchar, India for providing us the GPU-equipped workstations which were indispensable for this work. Also, the Office of Head of Department, Department of Computer Science and Engineering at National Institute of Technology, Silchar also provided one GPU equipped workstation for this work for which we are greatly obliged.

This work was not supported by any financial grant and there do not exist any conflicts of interest.

REFERENCES

[1] Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ”Imagenet: A large-scale hierarchical image database.” In 2009 IEEE conference on computer vision and pattern recognition, pp. 248-255. Ieee, 2009.

[2] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. ”Sequence to sequence learning with neural networks.” In Advances in neural information processing systems, pp. 3104-3112. 2014.

[3] Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. ”Learning phrase representations using RNN encoder-decoder for statistical machine translation.” arXiv preprint arXiv:1406.1078 (2014).

[4] Karpathy, Andrej, and Li Fei-Fei. ”Deep visual-semantic alignments for generating image descriptions.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3128-3137. 2015.

[5] Mao, Junhua, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. ”Deep captioning with multimodal recurrent neural networks (m-rnn).” arXiv preprint arXiv:1412.6632 (2014).

[6] Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. ”Show and tell: Lessons learned from the 2015 mscoco image captioning challenge.” IEEE transactions on pattern analysis and machine intelligence 39, no. 4 (2016): 652-663.

[7] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. ”Neural machine translation by jointly learning to align and translate.” arXiv preprint arXiv:1409.0473 (2014).

[8] Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. ”Show, attend and tell: Neural image caption generation with visual attention.” In International conference on machine learning, pp. 2048-2057. 2015.

[9] Bernardi, Raffaella, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. ”Automatic description generation from images: A survey of models, datasets, and evaluation measures.” Journal of Artificial Intelligence Research 55 (2016): 409-442.

[10] Hossain, MD Zakir, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. ”A comprehensive survey of deep learning for image captioning.” ACM Computing Surveys (CSUR) 51, no. 6 (2019): 1-36.

[11] LeCun, Yann, Bernhard Boser, John Denker, Donnie Henderson, R. Howard, Wayne Hubbard, and Lawrence Jackel. ”Handwritten digit recognition with a back-propagation network.” Advances in neural information processing systems 2 (1989): 396-404.

[12] Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang et al. ”Imagenet large scale visual recognition challenge.” International journal of computer vision 115, no. 3 (2015): 211-252.

[13] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. ”Deep residual learning for image recognition.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[14] Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[15] Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. ”Densely connected convolutional networks.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708. 2017.

[16] Kojima, Atsuhiro, Takeshi Tamura, and Kunio Fukunaga. ”Natural language description of human activities from video images based on concept hierarchy of actions.” International Journal of Computer Vision 50, no. 2 (2002): 171-184.

[17] Farhadi, Ali, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. ”Every picture tells a story: Generating sentences from images.” In European conference on computer vision, pp. 15-29. Springer, Berlin, Heidelberg, 2010.

[18] Mason, Rebecca, and Eugene Charniak. ”Nonparametric method for data-driven image captioning.” In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 592-598. 2014.

[19] Kulkarni, Girish, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. ”Babytalk: Understanding and generating simple image descriptions.” IEEE Transactions on Pattern Analysis and Machine Intelligence 35, no. 12 (2013): 2891-2903.

[20] Li, Siming, Girish Kulkarni, Tamara Berg, Alexander Berg, and Yejin Choi. ”Composing simple image descriptions using web-scale n-grams.” In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pp. 220-228. 2011.

[21] Elman, Jeffrey L. ”Finding structure in time.” Cognitive science 14, no. 2 (1990): 179-211.

[22] Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In NIPS, pp. 1097–1105, 2012.

[23] Hochreiter, Sepp, and Jürgen Schmidhuber. ”Long short-term memory.” Neural computation 9, no. 8 (1997): 1735-1780.

[24] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In arXiv:1502.03167, 2015

[25] Gehring, Jonas, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. ”Convolutional sequence to sequence learning.” arXiv preprint arXiv:1705.03122 (2017).

[26] Aneja, Jyoti, Aditya Deshpande, and Alexander G. Schwing. ”Convolutional image captioning.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5561-5570. 2018.

[27] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. ”Attention is all you need.” In Advances in neural information processing systems, pp. 5998-6008. 2017.

[28] Yu, Jun, Jing Li, Zhou Yu, and Qingming Huang. ”Multimodal transformer with multi-view visual representation for image captioning.” IEEE Transactions on Circuits and Systems for Video Technology (2019).

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. arXiv preprint arXiv:1409.4842, 2014.

[30] Young, Peter, Alice Lai, Micah Hodosh, and Julia Hockenmaier. ”From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions.” Transactions of the Association for Computational Linguistics 2 (2014): 67-78.

[31] Iandola, Forrest N., Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. ”SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and¡ 0.5 MB model size.” arXiv preprint arXiv:1602.07360 (2016).

[32] Ma, Ningning, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. ”Shufflenet v2: Practical guidelines for efficient cnn architecture design.” In Proceedings of the European conference on computer vision (ECCV), pp. 116-131. 2018.

[33] Sandler, Mark, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. ”Mobilenetv2: Inverted residuals and linear bottlenecks.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510-4520. 2018.

[34] Tan, Mingxing, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. ”Mnasnet: Platform-aware neural architecture search for mobile.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2820-2828. 2019.

[35] Zoph, Barret, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. ”Learning transferable architectures for scalable image recognition.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8697-8710. 2018.

[36] Chen, Yunpeng, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. ”Dual path networks.” In Advances in neural information processing systems, pp. 4467-4475. 2017.

[37] Xie, Saining, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. ”Aggregated residual transformations for deep neural networks.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1492-1500. 2017.

[38] Zagoruyko, Sergey, and Nikos Komodakis. ”Wide residual networks.” arXiv preprint arXiv:1605.07146 (2016).

[39] Hu, Jie, Li Shen, and Gang Sun. ”Squeeze-and-excitation networks.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132-7141. 2018.

[40] Zhang, Xingcheng, Zhizhong Li, Chen Change Loy, and Dahua Lin. ”Polynet: A pursuit of structural diversity in very deep networks.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 718-726. 2017.

[41] Szegedy, Christian, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. ”Inception-v4, inception-resnet and the impact of residual connections on learning.” arXiv preprint arXiv:1602.07261 (2016).

[42] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. ”Microsoft coco: Common objects in context.” In European conference on computer vision, pp. 740-755. Springer, Cham, 2014.

A Perceptual Matching based Deduplication Scheme using Gabor-ORB Filters for Medical Images

Sonal Ayyappan[1]
Reasearch Scholar
Department of Computer Science and Engineering
SRMIST, Chennai, India

Dr. C Lakshmi[2]
Professor and HOD
Department of Software Engineering
SRMIST, Chennai, India

***Abstract*—In the ever widening field of telemedicine, there is a greater need for intelligent methods to selectively choose data that are relevant enough to be transmitted over a network and checked remotely. By the very nature of medical imaging, a large amount of data is generated per imaging or scanning session. For instance, a Magnetic Resonance Images (MRI) scan consist of hundreds to thousands of images related to slices of the organ being scanned. But at often times all of these slices are not of interest during the process of medical diagnosis by the medical practitioner. Not only does this result in the access of unwanted data remotely, but it can also put greater strain on the bandwidth available over the network. If the relevant images can be selected automatically without human intervention, ensuring great sensitivity, the above-mentioned issues can also be alleviated. This paper proposes a novel method of perceptual matching and selection of relevant MRI images by using a deduplicating technique of combining Gabor filter with Oriented FAST and Rotated BRIEF (ORB) feature extraction technique on a vast set of MRI scan images. The outcome of this method are relevant deduplicated MRI scan images which can save the bandwidth and will be easy for the medical practitioner to verify remotely.**

***Keywords*—Perceptual matching; ORB feature extraction; Gabor filters; MRI scan; deduplication**

I. Introduction

Advanced data frameworks have been progressively conveyed in the recent medical care scenarios. Indeed, numerous medical clinics and hospitals depend on medical clinic data frameworks (HIS), radiology data frameworks (RIS), and picture documenting and correspondence frameworks (PACS), for storage of MRI Scan Medical images. These frameworks facilitate the practitioners to share images [1][2][]3]. Data Deduplication is a method of eliminating repeated copies of data in order to retain the storage capacity. It results in decreased cost per gigabyte with more area to store data. In radiology centers, more than petabytes of medical images are stored every year which may contain redundant data too. This increase of repeated data cannot be handled by the existing IT technologies [23]. In an MRI Scan collection of images for a single subject regarding any organ, views are taken from three directions. Each direction concentrates on taking images of slices of the organ with slight differences like 1mm, 4mm, 6mm etc. For a non suspicious subject, there may be more irrelevant repeated data which when stored will waste a lot of storage space. By the present nature of MRI scans of a single subject, relevant details are not distributed evenly in equidistant slides. Hence there is a need for perceptual matching and deduplication for transmitting medical images among practitioners for expert opinions[1]. By combining Gabor filter, ORB key-point detection and Brute-Force Matcher, we achieve this perceptual deduplication where relevant data is extracted without any specific regard for the distances between the slices. A finely tuned Gabor filter provides the ORB algorithm with just the right amount of details that leads to the most optimal matching differences.

This method, being based on ORB, has got excellent performance in identifying and matching near-similar images even if a one-to-one positional correspondence between the pixels are absent. Therefore, variations like missing regions, appreciable level of orientation difference, size difference, etc. can be accounted.

In this paper the method for fine tuning the parameters of Gabor Filter is implemented. It also describes the method of feature extraction using ORB and brute force matching technique which assures the performance of perceptual matching so as to identify repeated data and discard it. Section 2 give an overview of the work done. The rest of the paper is organized as follows. Section 3 explains the algorithms used for finding out the deduplicated images and the background where this method has its relevance. In Section 4, results are discussed. The paper is concluded in Section 5.

II. Literature Review

Storage used for medical imaging must be expandable in the coming years. It is normal for an incoming of Terabytes of medical imaging data per year for hospital radiology departments [4]. Deduplication is a very important aspect for the reliability of storage facilities, as the expense of storage management can be minimized by removing duplicated files [5]. A huge amount of memory is wasted in storing the redundant data for a single patient in a single take [24].

However, deciding whether two original files are the same or not by viewing two processed images is not easy. Message-locked encryption was suggested to facilitate the deduplication of processed image. This approach lets users produce the same ciphertext for the same file and enables users to reap the benefits of ciphertext deduplication [6]. Techniques had been proposed for the deduplication of encrypted data in different levels. To support the deduplication feature for partially duplicated files, block-level deduplication methods were studied [7]. Cloud media centers were implemented for dedicated deduplication system where the older versions were not working [8]. Near-duplicate data scanning for encrypted data has been studied. A lot of wastage in the storage system occurs due to these duplicates which is very costly to afford [9].

Deduplication techniques are characteristics based implemented. Characteristics like extracting features, using hash technique for indexing extracted image feature, image similarity detection based on distance using threshold, etc were used to find almost equal images. In order to analyze the exact duplicates or the near exact duplicates different features extraction algorithms like Scale Invariant Feature Transform, Speed Up Robust Feature, Principal Component Analysis, Binary Robust Independent Elementary Features were taken into account [10].

The authors in [11] used Difference of Gaussians, Principal Component Analysis with Scale Invariant Feature Transform methods for feature extraction of images and space efficient bloom filter were hashed with these features. The locality sensitive hashing used correlated attributes to find the similar images. This method gave efficient bandwidth and saved storage resources. Later in [12] the authors used perpetual hash algorithms for creating image signatures which identified similar duplicated images. It resulted in deduplicating storage and saving bandwidth while transmission. Another method used was MapReduce technique. It manages data in huge quantities in a distributed manner. The authors in [13] used this technique resulting in a faster image duplication identifier. SIFT methods where made strong by incorporating k-means algorithm and groups of multiple image clusters. This helped to detect the near duplicate images and those were hashed based on histogram distance [14]. Another research was with Local-based binary representation which made use of binary vector and histogram for finding out the duplicate images [15]. Later a real time novel method which made use of Bloom filters along with the existing techniques was implemented. It used the correlation property and resulted in reduced latency processing [16].

A faster and more efficient feature point detector than SIFT and SURF named Oriented FAST (Features from Accelerated Segment Test) and Rotated BRIEF(ORB) technique is nowadays used widely [17]. It has advantages of low computation cost and better performance [18]. Medical images can be exchanged in a very efficient way using the cloud [25]. It has the limitation of numerous applications pointing to the same data at the same time. Due to the increasing need of storage capacity, the PACS also have its own limitations [26]

III. Background

A complete MRI scan image set is retrieved from the MRI scanner and used by a medical practitioner for medical diagnosis. But in case an expert opinion is required from another medical practitioner, the files are first deduplicated to select only the relevant slides. This subset of the scan image set is then encrypted using a pixel-scrambling algorithm which uses chaotic maps and intensity variations[27]. On the receiver end, the files are decrypted to retrieve the deduplicated set of scan images. Due to the deduplication method applied, the bandwidth for the transferring of medical images among practitioners and the storage for storing the relevant data can be increased for some more time. In Fig. 1 the graphical representation of the above said scenario has been explained.

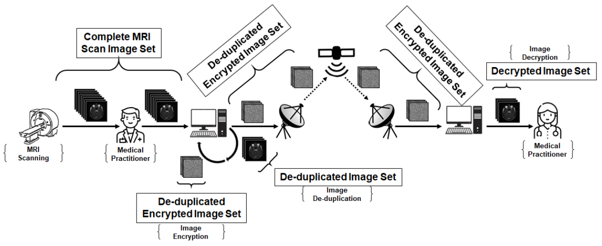


Fig. 1. Graphical Representation of the Use-Case Scenario of MRI Scan Image Deduplication.

A. Algorithms

Input images are chosen from complete MRI scan image set with very less slice width. Images 1 mm apart with no perceptible differences are chosen as similar images, and images 5 mm apart with some perceptible differences of valid nature are selected as dissimilar images for the execution of this method. Following are the algorithms used to find the most optimal Gabor filter parameter set. Gabor filter has parameters which needs to be fine tuned individually for this deduplication method.

```
Algorithm 1: Build_filters(image1, image2)
input  : ksize, sigma, theta, lambda, gamma
output: orb_values, Orb_match.csv
for ksize in [3,31] do
    for sigma in range (1,20), step-size: 1 do
        for theta in range (0, pi), step-size: pi/16 do
            for lambda in range(10,100), step-size: 10 do
                for gamma in range(0.1,1.2), step-size: 0.2 do
                    Initialize Gabor filter kernel with getGaborKernel(ksize, sigma, theta,lambda, gamma, 0) → kernel
                    Apply filter by calling filter2D(image1/image2,kernel) → fimage1, fimage2.
                    Call orb(fimage1, fimage2) → KeypointA, KeypointB, orb_match
                    Call Sklearn structural_similarity(fimage1,fimage2) → ssim
                    Append (Ksize, sigma, theta, lambda, gamma, KeypointA, KeypointB, ssim, orb_match) to orb_values list.
Create pandas dataframe orb_df from orb_values
Save orb_df as csv file Orb_match.csv
```

Algorithm 1 explains the procedure for fine tuning the parameter set for building the Gabor filter. Using these parameters Algorithm 3 sets the Gabor filter. Algorithm 2 provides the base condition for iterating ORB feature matching technique. The perpetual matching is carried out using Algorithm 4.

```
Algorithm 2: Base Parameter Iteration
input  : (image1, image2), (image2, image3)
output: orb_values, Orb_match.csv

Set default value for ksize, sigma, lambda, gamma
  from chosen_parameters.
Choose one parameter out of ksize, sigma, lambda,
  gamma to iterate → chosen_parameter.
for chosen_parameter in chosen_parameter_list do
    Build gaborfilter kernel using
      build$_{f}ilter(defaultparameters, chosen\_parameter) \rightarrow$
      filter_list
    Apply filter by calling
      process(image1/image2/image3, filter_list) →
      fimage1, fimage2, fimage3.
    Call orb(image1, image2) → SKeypointA,
      SKeypointB, Sorb_match.
    Call orb(image2, image3) → DKeypointA,
      DKeypointB, Dorb_match.
    Find match_difference = Sorb_match -
      Dorb_match
    Append default parameters, chosen parameter,
      (S/D)Keypoint(A/B), Sorb_match, Dorb_match
      to orb_values list.
Create pandas dataframe orb_df from orb_values
Save orb_df as csv file Orb_match.csv
```

```
Algorithm 3: build_filter(ksize, sigma, lambda,
gamma)
input  : ksize, sigma, lambda, gamma
output: filter_list

for theta in range (0, pi), step-size: pi/16 do
    Initialize Gabor filter kernel with
      getGaborKernel(ksize, sigma, theta, lambda,
      gamma, 0) → kernel
    Append kernel to filter_list
Return filter_list.
```

```
Algorithm 4: ORB Extraction and Brute-Force
Matching
input  : ORB Extraction and Brute-Force Matching
output: Matching_Percentage

Convert input images to gray-scale images → (img1,
  img2)
Create ORB object using ORB_create(nfeatures =
  1000, scoreType = FAST)
Detect key-points from input images, detect(img1,
  img2) → (kp1, kp2)
Compute ORB descriptors from input images and
  key-points → (kp_1, desc_1, kp_1, desc_2)
Create BFMatcher object with Hamming distance
  parameter.
Find matches between the images using descriptors,
  match(desc_1, desc_2) → matches[]
set lower = minimum number of key-points.
Matching_Percentage = sizeof(matches)/lower * 100
return Matching_Percentage.
```

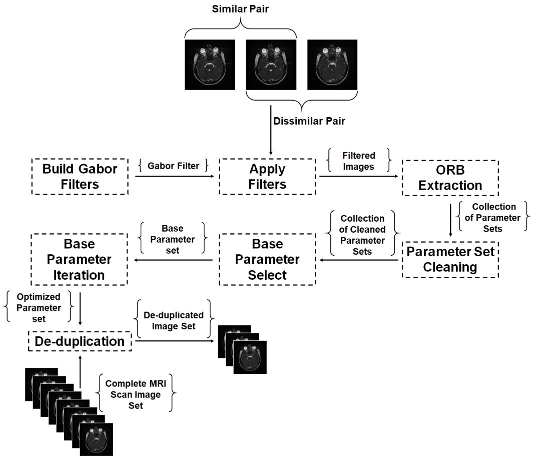


Fig. 2. Steps involved in the Proposed Deduplication Algorithm (1. Build Gabor Filters, 2. Apply Gabor Filters, 3. ORB Extraction, 4. Parameter Set Cleaning, 5. Base Parameter Select, 6. Base Parameter Iteration, 7. Deduplication

The resultant of these algorithms are a set of nearly matched MRI images which can be regarded deduplicated and can be discarded before transmitting.

B. Working

Building Gabor filters and its application, ORB feature extraction, parameter set cleaning, base parameter set selection and iteration, deduplication are the key phases in the proposed deduplication scheme. This scheme is graphically represented in the Fig. 2.

The prevailing section deals with each step of working of the whole deduplication procedure.

1) Build Gabor Filters: Gabor filters are special classes of band-pass filters which allow certain bands of frequency and reject the rest. The Gabor filters are well known for its time and frequency transform characteristics. Filters with distinct scaling directions can be constructed using Gabor filters caused by different parameters[19]. In image processing, they are used for feature extraction, edge detection, texture analysis, etc. The Gabor function has the capability to capture the localized information with respect to spatial frequency, location, and selection of direction [20]. Mathematically Gabor filters [21,22]are expressed as a function:

$$g(x, y; \lambda, \theta, \phi, \sigma, \gamma) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)\exp\left(i(2\pi\frac{x'}{\lambda}+\phi)\right)$$

$$x' = x\cos\theta + y\sin\theta$$
$$y' = -x\sin\theta + y\cos\theta$$

Build_filter function iterates through all Gabor filter parameters and create kernels for each set of parameters. Range of values for each parameter, namely, Ksize(x,y), Sigma(σ), Theta(θ), Lambda(λ) and gamma(γ) are predefined during this phase. These ranges are as follows: Ksize from the set [3,31], Sigma from 1 to 10 with a step size of 1, Theta from 0 to π with a step size of $\pi/16$, Lambda from 10 to 100 with a step size of 10 and Gamma from 0.1 to 1.2 with a step size of 0.2.

2) Apply Gabor Filters: Three images are selected to form two pairs of images, where one image is common to both. First pair is formed from MRI images that are taken 1 mm apart and represents similar images. Whereas second pair represents dissimilar images and are taken 5 mm apart. Pairs of similar and dissimilar images are read into the program and they are passed to the apply_filter function. Both images are filtered using the kernels and filtered Gabor images are formed.

3) ORB Key-point Extraction: These images are compared using ORB key-point extraction and brute-force matcher (BFMatcher). Orb function returns the number of key-points found in each image and matching percentage between them. Structural similarity index is also found between the images. These values along with Gabor filter parameters are saved into a CSV file for further processing.

4) Parameter Set Cleaning: The CSV file is cleaned so as to extract valid set of parameters. The cleaning process is as follows: any parameter set which gives very low number of key-points are discarded. All sets which generates very low or very high matching percentages are also discarded. Any set that produces abnormally high structural similarity index is discarded since high structural similarity index would indicate high loss of details in the images. A high minimum-threshold number of key-points in either images indicate existence of appreciable level of details.

5) Base Parameter Select: Parameter sets are sorted in the descending order of their matching percentages. First few parameter-sets with the highest matching percentages are retained and rest are discarded. These parameters are used to find matching percentages for dissimilar pair of images.

Difference between matching percentages of similar and dissimilar pairs are found for each parameter-set. These parameters along with key-point counts, matching percentages and matching percentage differences are stored as a CSV file. They are then sorted in the descending order of their matching percentage differences. And parameter set with the highest difference is selected as the base or default parameter set for further steps.

This parameter set represents the maximum gap between matching percentages of similar and dissimilar images. Therefore, it is the best provisional choice to decide whether two images are similar or not.

6) Base Parameter Iteration: Keeping the base parameter-set as the default values for each parameters, one parameter is chosen at a time to be iterated over a range of values which are characteristically valid for that particular parameter except for Theta. For every iteration, 16 Gabor filter kernels with Theta value ranging from 0 to π with a step size of $\pi/16$ are formed and each set of 16 images are combined to form the new pair. ORB key-point extraction and Brute-Force matching is performed on this new pair of images. Metrics like key-point counts, structural similarity, matching percentages and matching percentage differences are found and stored for further analysis.

Values of each parameter that gave the highest match differences in their respective iterations were selected and used to form a final parameter-set. This was further used to find the matching differences between similar and dissimilar images. This is done to check if parameter values that individually produced the best matching difference along with other default values could produce better results when collected together as a single parameter set. Depending on the result, it was either retained or discarded.

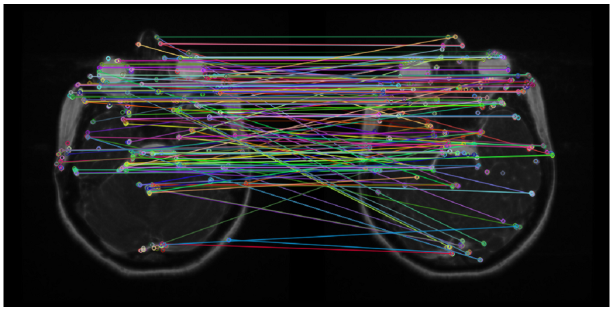


Fig. 3. Brute-Force Matching of ORB Key-Points on Filtered Images

7) Deduplication: After finding the best case parameter set for Gabor filter, it can now be used for deduplication on the entire MRI Scan Image Set. The filter now being fine-tuned for MRI scans, we can set well defined thresholds when two images are similar or dissimilar. By running this in an iterative manner, the entire set is deduplicated and only relevant slides or images are retained.

The result of finding out feature matching points using ORB feature extracting and Brute Force matcher is shown in Fig. 3.

IV. Result and Discussion

All the above mentioned steps were tried using SIFT, SURF and ORB key-point extractions and their preferred matchers. It was found that ORB provided the best-case efficiency in both matching performance and execution time. Hence ORB along with Brute-Force matcher was selected as the algorithm/technique to be used for further analysis in the paper.

Matching performance evaluation is done using the number of ORB key-point extracted and 'Match Differences'.

M = S - D Match Difference = S - D where, M = Match Difference S = Matching percentage between similar images D = Matching percentage between dissimilar images

Results of the iterated base parameter set for fine tuning each parameter of the Gabor filter is as follows.

A. Ksize

K-size is the size of Gabor-filter kernel. Here K-size of 31 represents a kernel of size 31x31 pixels. In order to improve time-efficiency of initial iterations, just a single pair of ksize values were selected. This pair (3,31) represents two possible extremes of values that k-size could take. In later iterations along with a variable k-size and default values for other parameters, it was noted that the key-point count and matching difference stabilized after k-size value of 13. Fig. 4 shows graphically this iterations. Hence the assumption was validated that time-efficiency was improved by choosing two extremes

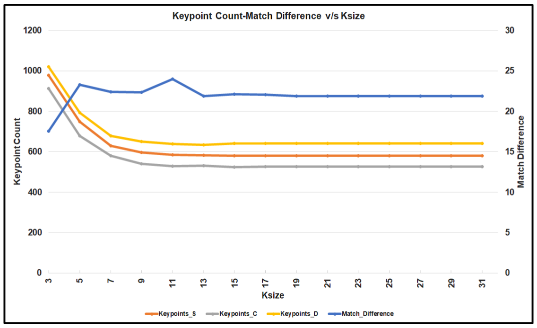


Fig. 4. Variation of Matching Difference and Key-Point Count vs the Iteration of Ksize

TABLE I. Match differences with variation of Ksize

ksize	KeyPtS	KeyPtC	KeyPtD	MatchDiff
3	978	912	1019	17.54386
5	748	678	792	23.30383
7	630	581	678	22.37522
9	597	541	650	22.36599
11	585	529	638	24.00756
13	583	530	635	21.88679
15	581	525	640	22.09524
17	580	526	640	22.05323
19	580	526	640	21.86312
21	580	526	640	21.86312
23	580	526	640	21.86312
25	580	526	640	21.86312
27	580	526	640	21.86312
29	580	526	640	21.86312
31	580	526	640	21.86312

of values rather than an extensive range. It also shows that matching difference is sensitive to k-size individually for a limited range values and the default k-size value that gave the best overall performance could very well be outside the high sensitivity range.

Table I gives a summary of the similar keypoint count, dissimilar keypoint count, total matched keypoints and match differences with varying ksize in the Gabor filter. The iteration was done with an increase of 2 starting from 3 to 31. Above this, the results were repeating and constant.

B. Sigma

Sigma represents the bandwidth of the Gabor envelop. A higher sigma values increases the overall size of the envelop and allows for more strips (frequency bands) through it.

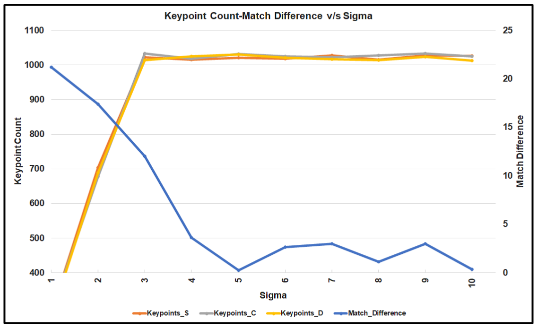


Fig. 5. Variation of Matching Difference and Key-Point Count vs the Iteration of Sigma

TABLE II. Match differences with variation of Sigma

Sigma	KeyPtS	KeyPtC	KeyPtD	MatchDiff
1	580	526	640	21.86312
2	285	290	274	21.20246
3	704	678	689	17.40413
4	1022	1033	1014	12.00087
5	1015	1018	1025	3.641739
6	1021	1032	1030	0.227647
7	1018	1025	1021	2.631381
8	1028	1022	1017	3.002192
9	1016	1028	1014	1.121504
10	1028	1033	1023	2.980191

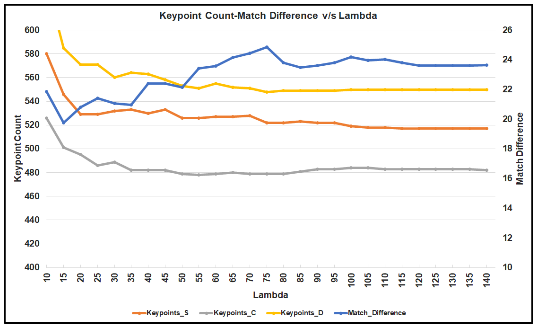


Fig. 6. Variation of Matching Difference and Key-Point Count vs the Iteration of Lambda

It was noted that the matching difference for medical scan images are highest for low values for sigma. This is despite a sudden increase in the key-point count as is shown from the graph. This can be clearly seen in Fig. 5. As can be noted, key-point count stabilizes after sigma of 3, but matching difference has got an overall downward trend and keep fluctuating at higher values of sigma. This is due to larger amount of details that is passed through the Gabor filter for higher sigma values This results in adversely affecting matching performance even though there are more features to extract key-points from. Hence matching performance is highly sensitive to sigma. Sigma variation also has the greatest effect on perceptual state of the output image. Table II shows the iterated values for fixing the sigma value.

C. Lambda

Lambda or wavelength governs the width of the Gabor function. Higher lambda values produce thicker Gabor strips. As can be noted from Fig. 6, there is appreciable improvement in matching difference as lambda increases till a specific level. Meaning increasing amount of details in the image due to wavelength change is constructive towards better matching difference performance. But beyond the point of optimality, more details work against the match and hence do not produce the best possible matching difference. But overall stability of matching difference is quite high for any higher values for lambda after the point of optimality. Table III shows the iterated values for stabilizing the lambda values

D. Gamma

Gamma determines the aspect ratio of the Gabor filter. And it controls the height of the function. A very high

TABLE III. MATCH DIFFERENCES WITH VARIATION OF LAMBDA

lambda	KeyPtS	KeyPtC	KeyPtD	MatchDiff
10	580	526	640	21.86311787
15	546	501	585	19.76047904
20	529	495	571	20.80808081
25	529	486	571	21.39917695
30	532	489	560	21.06339468
35	533	482	564	20.95435685
40	530	482	563	22.406639
45	533	482	558	22.406639
50	526	479	553	22.12943633
55	526	478	551	23.43096234
60	527	479	555	23.5908142
65	527	480	552	24.16666667
70	528	479	551	24.42588727
75	522	479	548	24.8434238
80	522	479	549	23.79958246
85	523	481	549	23.49272349
90	522	483	549	23.60248447
95	522	483	549	23.80952381
100	519	484	550	24.17355372
105	518	484	550	23.96694215
110	518	483	550	24.01656315
115	517	483	550	23.80952381
120	517	483	550	23.60248447
125	517	483	550	23.60248447
130	517	483	550	23.60248447

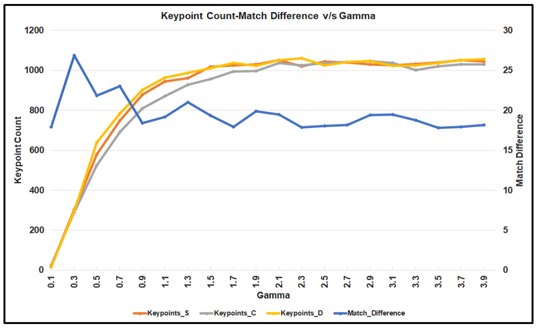


Fig. 7. Variation of Matching Difference and Key-Point Count vs the Iteration of Gamma

gamma results in lower height of the filter, while a lower value increases the height. For a series of increasing gamma values, a seesaw behavior in the matching difference was observed. For lower number of key-point counts, the matching difference appreciated with an increasing gamma, while at higher ranges with more details in the images, it showed depreciation followed by stability. Table IV shows the iteration for finding the gamma value for Gabor filter and is graphically shown in Fig. 7.

E. Discussion

MRI Scan Images of a subject was taken at two different periods live from a radiology center. The subject under experiment was suffering from brain tumor while diagnosed. This paper deals with a method where the subject or the patient has undergone a treatment and it on regular follow-ups. Thus, this methods from all the above observations apparently justifies that MRI scan images of this particular type works very well within a narrow set of values of filter parameters. And hence it validates the need for an algorithm to fine tune filter parameters to the application, as is proposed in this paper. When each parameter value from their individual point of optimality was clubbed together to determine the matching difference, it was noted that it faired considerably inferior to the initial base parameter set. Hence it is clear that the matching performance is sensitive to all the parameters as a whole. The base parameter-set obtained provided the best performance and could be used for deduplication of medical scan images of this particular kind reliably well. Bench mark datasets were not used as it didn't provide a whole set of MRI images for a particular subject and at two or more different periods for the same subject.

TABLE IV. MATCH DIFFERENCES WITH VARIATION OF GAMMA

Gamma	KeyPtS	KeyPtC	KeyPtD	MatchDiff
0.1	22	26	17	17.9144385
0.3	302	290	296	2689655172
0.5	580	526	640	21.86311787
0.7	749	691	783	23.01013025
0.9	879	809	901	18.41779975
1.1	944	871	963	19.17336395
1.3	962	928	988	21.01293103
1.5	1017	957	1011	19.33124347
1.7	1025	994	1038	17.90744467
1.9	1031	997	1022	19.85957874
2.1	1052	1038	1051	19.46050096
2.3	1021	1026	1060	17.86375014
2.5	1043	1037	1024	18.07817774
2.7	1039	1042	1042	18.17329523
2.9	1031	1047	1046	19.39734391
3.1	1026	1037	1022	19.48650164
3.3	1032	1002	1025	18.76247505
3.5	1039	1021	1038	17.82566112
3.7	1052	1030	1052	17.96116505
3.9	1045	1029	1056	18.17298348

V. CONCLUSION

In the real-world conditions where telemedicine is getting more prevalent day by day, MRI scan images needs to be transmitted over open network. This puts a considerable strain on the bandwidth and security requirement since the data is considerably large in size and highly confidential in nature. Hence by using the deduplication scheme suggested in this paper, the number of scan images can be reduced by selecting a subset of almost similar scan images. This subset would only contain scan images which contain relevant details and hence redundancy is avoided. Using chaotic maps that ensure security over data transmission, each image can be encrypted. Thus tackling two of the major challenges of telemedicine successfully.

REFERENCES

[1] J. Fiaidhi, C. Kuziemsky, S. Mohammed, J. Weber, T. Topaloglou. *Emerging IT trends in healthcare and well-being*, 18th ed, vol 3, pp 9-13, IT Prof, 2016

[2] N.Saranummi *In the spotlight: health information systems* 6th ed, pp 21- 23, IEEE Rev Biomed Eng, 2013.

[3] H.K Huang *PACS and imaging informatics: basic principles and applications* New York: Wiley-Blackwell,2010.

[4] *https://searchhealthit.techtarget.com*

[5] M W Storer, K Greenan, D D E Long, E L Miller, *Secure data deduplication* pp 1 - 10, 4th ACM International Workshop on Storage Security and Survivability; ACM: New York, NY, USA, 2008.

[6] M Bellare, S Keelveedhi,T Ristenpart, *Message-locked encryption and secure deduplication*, Annual International Conference on the Theory and Applications of Cryptographic Techniques; Springer: Berlin/Heidelberg, Germany, 2013.

[7] R Chen, Y Mu, G Yang, F Guo, *BL-MLE: Block-Level Message-Locked Encryption for Secure Large File Deduplication*, vol 10, pp 2643 - 2652, IEEE Trans. Inf. Forensics Security, 2015.

[8] Y Zheng, X Yuan, X Wang, J Jiang, C Wang, X Gui, *Toward Encrypted Cloud Media Center with Secure Deduplication*,vol 19, pp 251 - 265, IEEE Trans. Multimedia, 2017.

[9] H Cui, X Yuan, Y Zheng, C Wang, *Enabling secure and effective near-duplicate detection over encrypted in-network storage*, pp 1 - 9, IEEE INFOCOM 2016—The 35th Annual IEEE International Conference on Computer Communications; IEEE: New York, NY, USA, 2016.

[10] R Kaur , I Chana , J Bhattacharya *Data Deduplication Techniques for efficient cloud storage management: a systematic review*, vol 74, pp 2035 -2085, Journal of SuperComputing, 2018.

[11] Y Hua , W He, X Liu , D Feng *SmartEye: real-time and efficient cloud image sharing for disaster environments*, pp 1616 -1624, IEEE Conference on Computer Communications, 2015.

[12] X Li, J Li, F Huang *A secure cloud storage system supporting privacy-preserving fuzzy deduplication* vol 20(4), pp 1437 - 1448, Journal of Soft Computing, 2016.

[13] A S Deshmukh, P D Lambhate *A methodological survey on mapreduce for identification of duplicate images*, vol 5(1), pp 2016 - 210, International Journal of Scientific Research, 2016.

[14] Z Li, X Feng. *Near duplicate image detecting algorithm based on bag of visual word model*, vol 8(5), pp 557-565, Journal of Multimedia, 2013.

[15] F Nian ,T Li , X Wu , Q Gao , F Li. *Efficient near-duplicate image detection with a local-based binary representation*,vol 75(5), pp 2435 - 2452, Multimedia Tools Applications, 2016.

[16] Y Hua , H Jiang , D Feng. *FAST: Near real-time searchable data analytics for the cloud*,pp 754 -765,IEEE Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2014.

[17] E. Rublee, V. Rabaud, K. Konolige and G. Bradski. *ORB: an efficient alternative to SIFT or SURF*,pp 2564 -2571 International Conference on Computer Vision, 2012.

[18] T Miroslav , H Mark. *Fast corner detection*, pp 75 -87, Image and Vision computing, 1998.

[19] J. G. Daugman. *Two-dimensional spectral analysis of cortical receptive field profiles*, vol 20(10), pp 847 -856,Vision Research, 1980.

[20] Y Zhang, W Li, L Zhang, X Ning,L Sun, Y Lu. *Adaptive learning Gabor filter for finger-vein recognition*, vol 7, pp 159821-159830, IEEE Access, 2019.

[21] J. G. Daugman. *Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression*, vol 36(7), pp 1169-1179,IEEE Transaction in acoustics speech & signal processing, 1988.

[22] M. Porat, Y. Y. Zeevi. *The generalized Gabor scheme of image representation in biological and machine vision*, vol 10(4), pp 452-468, IEEE Transaction in Pattern Analysis and Machine Intelligence, 1988.

[23] R. Glicksman. *Archiving: Fundamentals of Storage Technology*, Society for Imaging Informatics in Medicine.

[24] D.R.VarmaM. *Managing DICOM Images:Tips and Tricks for the radiologist*, vol 22(1), pp 4-13, Indian Journal of Radiology and Imaging, 2012.

[25] S.G.Shiny, T. Thomas, K. Chithraranjan *Cloud Based Medical Image Exchange Security Challanges*, vol 38, pp 3453-3461, Procedia Engineering, 2012.

[26] M.Alhajeri, S.G.S.Shah. *Limitations in and Solutions for Improving the Functionality of Picture Archiving and Communication System:an Exploratory Study of PACS Professionals' Perspective*, vol 32(1), pp 54-67, Journal of Digital Imaging, 2018.

[27] S. Ayyappan, C. Lakshmi. *Empirical Analysis of Robust Chaotic Maps for Image Encryption*, vol 9(11), pp 393-398, International Journal of Innovative Technology and Exploring Engineering, 2020.

Post Classification in the Social Networks using the Map-reduce Algorithm

Abdoulaye SERE[1], José Arthur OUEDRAOGO[2]
Network of Computer Science Teachers
and Scientists in Faso
Bobo-Dioulasso, Burkina Faso

Boureima ZERBO[3], Oumarou SIE[4]
Network of Computer Science Teachers
and Scientists in Faso
Ouagadougou, Burkina Faso

***Abstract*—Wrongdoing is increasing through social media. Detecting them requires highlighting the most interesting topics in the posts. This essential part in the characterization of social network users could be done by a classification of posts. For this, we use a tuple of keywords and the Map-reduce algorithm for data collection and extraction. The main purpose is to achieve success on software realization which will establish a network between social networks to extract data and to speed up the classification of posts. The proposed method consists of verifying a sequence of keywords in the posts, following a grammar in order to determine classes. It allows the categorization of posts and monitoring of social networks. The categorization facilitates research of a particular post containing specific words. Thus, we contribute to increase capacity for wrongdoing prevention and strengthening cyber-security.**

***Keywords*—Big Data; map-reduce; social network; cyber-security; classification**

I. Introduction

Now-a-days, data is playing essential roles in analysis for the enterprises, in taking decisions. Data has become the basic resources for all the applications. Particularly, the applications based on the techniques of deep Learning and Data mining need more data.

Inversely, the internet of Things, the social networks produce enormous quantity of data, named Big Data. Due to the number of posts, the social networks propose a diversity of Big Data with several types such as texts, images and video.

Big Data involves several levels of problems such as the problem of data collection, data processing and data visualization. The problem of Big Data visualisation has been analyzed by Alexandre Perrot in his thesis in [1].

Data extraction is essential to build the collections of data on infrastructures for analysis. Frederic and others in [3] used Netvizz to extract data from Facebook to realize surveys on the 2015 presidential election in Burkina Faso.

Thus, Big Data processing needs the parallelization of data and tasks in order to reduce processing time. Since 2004, Dean and others in [2] proposed the map-reduce framework to count words in enormous documents. The techniques of processing big data have increased more and more with several implementations such as Hadoop , Spark in [5] and in [7]. The map-reduce has been also used in [6] for data-mining and in [4] for parallel algorithmic . A survey on the map-reduce framework has been proposed by N.Alamelu Menaka and others in [10].

Both the map-reduce algorithm and classification are the subjects of several works in [8], [9]. For example, in [8] Ouatik and others have tried a classification of student into four classes, scientific, literary, technical and original, in using the map-reduce algorithm.

Hadoop has been improved to take image processing into account. For example, Hadoop Image Processing Image (HIPI) has been proposed by the University of Virginia Computer Graphics Lab, in 2016 : it increases Hadoop capacity with new functions, giving the processing of distributed images and in taking the techniques of image processing into account. SERE and others in [11] introduced an application of the Hough Transform based on the map-reduce algorithm to improve processing time in straight line detection in distributed big images. The works of SERE and others in [11] have been extended by Mateus Coelho and others in [12], to circle detection in using the map-reduce algorithm.

In others ways, in recent years, scientists have worked on multi-label classification to obtain web page categorization. In [13], Yaya and others have studied the multi label classification in using an ontology in order to classify the web pages.

Every day, users post regularly relevant informations on social networks that can be used to characterize them to determine their behaviour. The posts give a way to obtain information on the users, to predict their future behaviour, to profiling them and to know what they are doing or what they are planning to do.

A selection of keywords allows generally the detection of posts related to specific subjects such as wrongdoing, fake news, terrorism, covid 19. For example, a list of keywords in the posts leads to track the user behaviour.

A Hub between social networks implementing the map-reduce algorithm, using a set of keywords, to extract speedly data from social networks does not exist.

But, there are available interface layers such as Application Programming Interface (API) to extract posts from social networks using keywords. For instance, Twitter proposes an API in python named tweepy that allows to get all the posts extracted into a file.

In this paper, we address the issue of post categorization in social networks. The objective is to facilitate profiling by classifying the most discussed topics in social networks. To do them, the map-reduce algorithm to establish data parallelism is used. Also, in big data, to speed up data processing, we use parallelism in data extraction. The purpose focuses on monitoring of social networks through a software connected simultaneously on several networks.

The paper is organized as follows: the Section 2 concerns preliminaries followed by the method description in the Section 3. At the end, the Section 4 deals with the simulation of both parallelism using the library thread and parallelism through the map-reduce framework.

II. Preliminaries

This section is focusing on the concepts related to the theories of words, languages and explains the meaning of the map-reduce algorithm. That allows forward the best understanding of the following sections.

In the theories of words and languages, a word is a sequence of elements or symbols in the alphabet while a language is a finite set of words. A language can be defined by a regular expression or by a grammar. For instance, let A be an alphabet {a, b, c}. "aa", "abc" are the words of the alphabet A. A^* is the universal language. It is the set of all the words defined by the alphabet A. Any word based on the alphabet A is a member of A^*. Any language based on the alphabet A is a subset of A^*.

In this paper, the keywords could be the words with any alphabet for any language such as French or English used to write posts.

The map-reduce algorithm is proposed initially by Dean and others, in 2004 [2], when working at Google. It is made up of three phases: the mapping phase, the shuffle phase, the reducing phase. The map function defines a transformation that accepts as input a single key value (k, v) pair and produces as output a set of intermediate (key, value) pairs (k_i, v_i). At the end of all the map functions, several pairs (k_i, v_i) are produced for the shuffle phase. For instance the map function transforms the pair (k, v) as input and produces the set of pairs (k_i, v_i) such as (k_1, v_1), (k_2, v_2).

The shuffle phase defines the groups of pairs that have the same k_i. It plays the role of creating groups. Suppose that (k_1, v_1), (k_1, v_2), (k_1, v_3), (k_2, v_4), (k_2, v_5), (k_3, v_3) the pairs produced by all the map functions. The shuffle creates three groups where each group along with its key k_i such as summarized in the Table I:

- the group 1 is $(k_1, < v_1, v_2, v_3 >)$
- the group 2 corresponds to $(k_2, < v_4, v_5 >)$
- the group 3 is (k_3, v_3).

TABLE I. Three groups created by the shuffle phase

Groups	Data
Group 1	$(k_1, < v_1, v_2, v_3 >)$
Group 2	$(k_2, < v_4, v_5 >)$
Group 3	(k_3, v_3)

Forward in the method description, the value v_i will correspond to a post. k_i will mean the common criteria for all the posts in a class.

The reduce phase concerns mainly the processing of each group, created by the shuffle phase. That means an operation is introduced by the reduce function to perform the value v_i of each group. For example, considering the operation +, the values of previous groups become:

- the group 1 $(k_1, < v_1, v_2, v_3 >) \Longrightarrow (k_1, < v_1 + v_2 + v_3 >)$
- the group 2 $(k_2, < v_4, v_5 >) \Longrightarrow (k_2, < v_4 + v_5 >)$
- the group 3 $(k_3, v_3) \Longrightarrow (k_3, v_3)$

Finally, the results of the map-reduce algorithm give $(k_1, < v_1 + v_2 + v_3 >)$, $(k_2, < v_4 + v_5 >)$ and (k_3, v_3)

There is a similarity between the wordcount (Dean and others in 2004 [2]) and classification proposed in our method.

III. Method Description

This section describes the technique of data extraction used through API layer connected on social networks and the classification of posts based on the map-reduce algorithm.

In the social networks, a post is made of texts, images and video. Often, a text comments an image or a video. That means, the text analysis is useful for the descriptions of an image or a video. A text is a set of sentences or a sequence of words. For instance, the following message posted by the "Ministère du Développement de l'Economie Numérique et des Postes" in French in Facebook presents two sentences such as the sentence 1 and the sentence 2:

- the post is: "Les panélistes du webinaire de dissémination des stratégies de développement du numérique et des postes au Burkina Faso des 25 et 26 juin 2020. Le lien pour l'inscription: https://bit.ly/3gRDsy0";
- the sentence 1 is: "Les panélistes du webinaire de dissémination des stratégies de développement du numérique et des postes au Burkina Faso des 25 et 26 juin 2020." ;
- the sentence 2 is: "Le lien pour l'inscription: https://bit.ly/3gRDsy0".

A sentence is a sequence of words. Inversely, a sequence of words is not always a right sentence. Formally, a sequence of words is represented by a tuple of keywords defined by (w_1, w_2,..., w_{n-1}, w_n) where w_i is a keyword. In the previous post, a tuple (w_1, w_2,..., w_{n-1}, w_n) can be (panélistes , stratégie, développement).

Data extraction is based on the definition of a criteria.

A. Data Extraction

Data extraction concerns extraction of posts regularly in the social networks in following the keywords and to store them in a nosql database, in the perspectives to prepare the research phase on the posts. The problem is also how to determine the manner to realize Data extraction and collection, speedly.

Social networks propose facilities through API to access to the posts with certain constraints. The problems of Data extraction depends on the permissions of social networks and API layer availability to extract posts. The technique used for extraction of posts depends on each social network.

There are no connections between social networks through a common hub to facilitate extraction and the collection of posts in using keywords. That leads to establish a layer to realize data extractions.

Due to the large volume of posts from social networks, it appears necessary to use appropriate techniques in the layer such as the parallelism of data or tasks to process all the posts. For instance, in Twitter an Application Programming Interface (API) allows extraction of posts in using keywords. A solution is to establish the parallelization of available functions in this API to accelerate data extraction.

Thus, the proposed method introduces a layer named the map-reduce algorithm that will play the role of data extraction and will transfer data to a nosql database for data collection. It consists of injecting codes into the function map and reduce to define criteria for data extraction. The layer will communicate with the API layer in order to access to the social network database of posts. For instance, the Fig. 1 illustrates the global architecture of data extraction and collection.

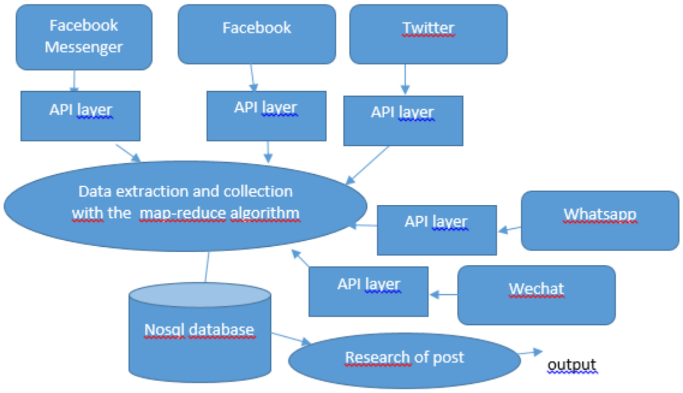


Fig. 1. The Architecture of Data Extraction, Collection and Research

B. Data Classification

In classification, a criteria establishes conditional instructions that describe how to put together data in the same group : It gives the common characteristics or the properties of data or the conditions that data must respect to be considered in the same class. The criteria are based on the tuple (w_1, w_2,..., w_{n-1}, w_n) and the grammar, defined by:

$$A1 : -w_1||w_1A2||A2$$
$$A2 : -w_2||w_2A3||A3$$
$$\overline{\qquad\qquad\qquad\qquad}$$
$$An-1 : -w_{n-1}||w_{n-1}An||An$$
$$An : -w_n||\epsilon$$

the grammar leads to respect the ordering of keywords, to align keywords syntactically. But ϵ is not considered as a keyword in this context. The method consists of the verification of a sequence of keywords appeared in the posts. The sequence of keywords produced by the criteria based on the tuple (w_1, w_2,..., w_{n-1}, w_n). The grammar follows the forms (w_i,... w_j) where i < j. The tuple defines the keywords and establishes the order of keywords that is confirmed by the grammar. The grammar also generates the groups of keywords. For instance, four groups of keywords w_1, (w_1, w_4, w_5), w_2, (w_4, w_{10}) will correspond respectively to four classes such as summarized in the Table II, which presents a similarity with the wordcount [2], corresponding to the Table I in preliminaries:

- the class 1 is the set P1 of posts that match the keyword w_1;
- the class 2 concerns the set P2 of posts that match the sequence (w_1, w_4, w_5) in the order;
- the class 3 is the set P3 of posts that match the keyword w_2;
- the class 4 concerns the set P4 of posts that match the sequence(w_4, w_{10}) in the order.

TABLE II. Table of four classes

Post classes	Keywords	Posts	Data
class 1	w_1	P1	(w_1, P1)
class 2	(w_1, w_4, w_5)	P2	((w_1, w_4, w_5), P2)
class 3	w_2	P3	(w_2, P3)
class 4	(w_4, w_{10})	P4	((w_4, w_{10}), P4)

If two posts contains the same sequence of words, they will belong to the same class. Thus, the order of words in the tuples influes on the result of classification. The order of keywords (w_i,... w_j) where i > j, is not considered in this analysis and could be studied in the perspectives.

In combinations of words, the number of combination for the keywords in the tuple (w_1, w_2,..., w_{n-1}, w_n) is established by $C_n^1 + C_n^2 + ... + C_n^{n-1} + C_n^n$ and corresponds to the number of post classes.

The ordering of the keywords, in following the tuple, determines a semantic meaning associated to a post that matches these keywords. For instance, a post can be a member of the class w_1 and the class w_2, if it matches the pair (w_2, w_1), not ordered in following the tuple. In this manner, a post can become a member of two different classes.

The classification is realized by the map-reduce algorithm, introduced here as a template to implement the parallelization of data and tasks and to accelerate data processing. By analogy to the works of Dean and others on the wordcount, to count the number of words in a document [2], the proposed model performs the tuple of keywords to verify if they appears in the order in the posts, to output the classes of posts. For instance, the algorithms 1, 2 show respectively the content of the map function and the reduce function that deal with the parallelization of data and tasks. The grammar is implemented in the content of the map function. The operation of classification is automatically done by the shuffle phase. The reduce function creates as outputs the classes.

In the Algorithm 1, tab[1..N] is a vector of keywords where each element corresponds to a class: The keywords generated by the grammar are inserted into the vector tab[1..N]. That means N= $C_n^1 + C_n^2 + ... + C_n^{n-1} + C_n^n$

```
Algorithm 1: function map(k:PostID, v:Post)
  Result: the pairs in the form $(k, v)$
  tab[1..N]: vector $< Keyword >$;
  p: Keyword ;
  i: Integer ;
  for (i=1; $i \leq n$; i++) do
      p=tab[i] ;
      if match(v, p) then
          emit(p, v);
      end
  end
```

Suppose that the shuffle transforms the set of pairs in the form (k, v) to the pairs $(k, < v_1, v_2, ...v_{k-1}, v_k >)$ where $< v_1, v_2, ...v_{k-1}, v_k >$, having a type vector $< Post >$, in order to prepare the reducing phase. As illustrated in the algorithm 2, the reducing phase removes duplicate posts in a vector of posts.

```
Algorithm 2: function reduce(p:pattern, list: vector
  $< Post >$)
  Result: the pairs in the form $(k, v)$
  cl: vector $< Post >$;
  index :vector $< int >$;
  temp: Post ;
  b:boolean ;
  i, j : Integer ;
  b=false;
  for (i=1; $i \leq n$; i++) do
      temp=list[i] ;
      for (j=i+1; $j \leq n$; j++) do
          if same(list[j],temp) then
              b=true;
          end
      end
      if (b==false) then
          cl.add(temp);
      end
      b=false;
  end
  emit (p, cl)
```

Simulations on the proposed algorithms will be studied in the Section IV.

IV. Numerical Simulation and Time Evaluation

This section evaluates the time according to the number of posts extracted on social networks.

Consider the tuple (w_1, w_2,..., w_{n-1}, w_n) where n=2. That corresponds to the tuple (w_1, w_2). Suppose that (w_1, w_2)=(terrorism, covid19). The table tab[1..N] of keywords in the Algorithm 1 become tab[1..3], represented by the Table III.

TABLE III. Three keywords

Keyword 1	Keyword 2	Keyword 3
terrorism	covid19	(terrorism, covid19)

TABLE IV. Three classes

Classes	Keywords
Class 1	terrorism
Class 2	covid19
Class 3	(terrorism,covid19)

The number of classes is then N= $C_2^1 + C_2^2$. That leads to three classes such as summarized in the Table IV.

The classification of posts is established by the shuffle phase in considering the keywords in the vector tab[1..3].

Global time processing depends on the configuration of both Local Area Network (LAN) and Wide Area Network (WAN). For LAN, data extraction uses a computer, connected on the Internet directly or through a proxy in a LAN. The characteristics of this computer are defined by :

- Memory : 3,7 GB
- Processor : Intel® Celeron(R) CPU B830 @ 1.80GHz x 2
- Graphic card : Intel® HD Graphics 2000 (SNB GT1)
- Operating System : Ubuntu 18,04 LTS 64 bits

In considering the pair of keywords (terrorism,covid19), a layer, written in python, is used to extract data from Twitter and Facebook, implementing data parallelism: The layer has been connected to Twitter server and Facebook server in using API, respectively tweepy and Facebook.

Here, parallelism implementations have been performed by two alternatives, such as data parallelism based on the library Thread in python and data parallelism implementation through the map-reduce algorithm with spark streaming.

A. Data Parallelism using the Library Thread (in Python)

In this way, three threads such as task 1, task 2, task 3 have been created. The task 1 is connected on Twitter to search the posts that correspond to the keyword “Covid19”. The task 2 also connected on Twitter, searches the keyword “terrorism” while the tack 3 performs the posts of Facebook. Thus, a common platform is created between Twitter and Facebook, to extract the posts. The Fig. 2 shows step by step according to the number of posts, processing time between the serial case and the parallel case. That also shows clearly, according to the number of posts extracted increasing, an improvement of processing time in the parallel case.

B. Data Parallelism using Spark Streaming

There exist many tools that could be used for data parallelism on real time. For instance, Apache storm [14], Apache Flink [15] and spark streaming [16] implement the framework map-reduce and can be connected directly to Twitter, as a data source, to obtain the posts. But here, the map-reduce in spark streaming, is connected on Twitter that has been implemented,

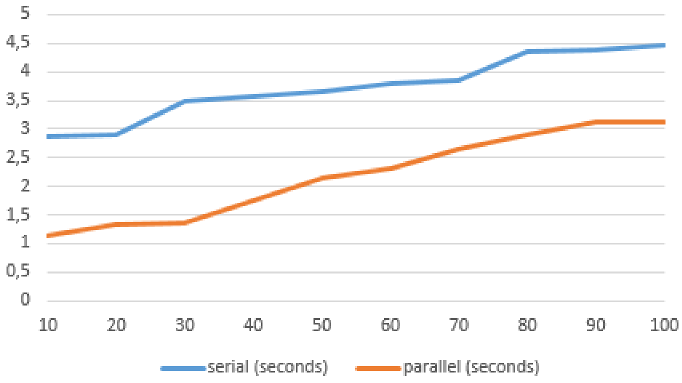


Fig. 2. Curves (Serial, Parallel)

in following the previous Algorithms 1 and 2: the keywords used in the function mapreduce are “terrorism”, “Covid19”. Processing time in the three cases (serial, parallel and spark streaming) are illustrated in the Fig. 3. That shows:

- from 0 to 70 tweets, tweet collection time in spark streaming is inferior to the serial case one and superior to the parallel case one.
- from 70 tweets to more, collection time in spark streaming is decreasing on the curve
- from 80 collected tweets to 100 collected tweets, the curve for spark streaming is going decreasing and stays under both the curve of the serial case and the curve of the parallel case.

Finally, the experimental results shows that data parallelism using a map-reduce tool in streaming, gives more posts in a short time.

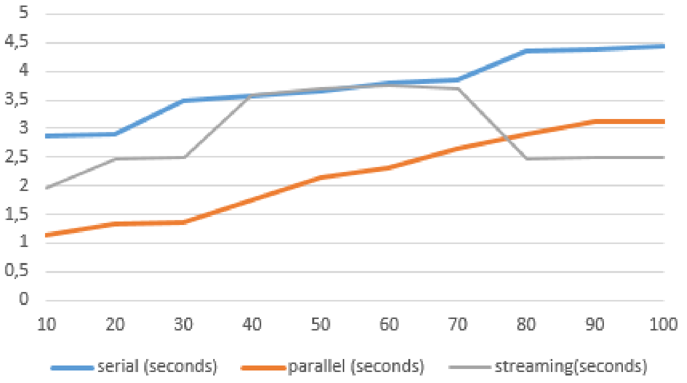


Fig. 3. All the Curves (Serial, Parallel, Spark Streaming)

Thus, the map-reduce tool, spark streaming is better than the serial case and the parallel case, in post extraction in the term of processing time, in considering the number of posts increasing up to 80.

The future research will focus on Deep analysis of network frames produced by spark streaming. We’ll study in details the content and the functionalities of spark streaming, noticeably the number of dynamic threads generated for scalability control according to the number of post increasing. It contributes to the best understanding of advantages with spark streaming case than in using directly the library thread.

Moreover spark streaming extension or others tools to take others social networks will lead easily to a software for monitoring social networks about post contents. Because, many social networks don’t accept unfortunately connection to the posts in streaming status through a map-reduce tool.

V. Conclusion

This paper has proposed the map-reduce algorithm, to extract speedly posts from social networks such as Twitter and Facebook, in order to realize a classification of posts, based on a tuple of keywords and a grammar. The tuples has defined the order of keywords while a grammar has generated the groups of keywords, corresponding to the classes. The keywords also play the role of class indexation.

The experimental resuls reveal that spark streaming is better than the parallelism using the library thread (in python) in post extraction in the term of processing time, in considering the number of posts to extract increasing, for instance superior to 80 posts in our study.

In perspectives, data extraction will be extended to more social network to have more posts, according to API layer availability to connect on social network. Through spark streaming for instance, it will be interesting to introduce more social network accessibility to the posts, not only with Twitter, but with proposed API Layer for each social network in order to strengthen the common platform as the main purpose.

The future works will also focus on content analysis of spark streaming functionalities and the study of class indexation, to facilitate research on stored posts to highlight data characteristics.

The tuples of keywords (w_1, w_2,..., w_{n-1}, w_n) could be the parameters for the tuples of ontologies in a specific domain. That lead to others deep analysis with the map-reduce algorithm, in perspectives.

Acknowledgment

The authors would like to thank the Ministry of Numeric Economy and Post (MDENP) for finding relevance in using big data opportunities for digital transformation of public administrations through several events such as FNDSI 2020 (Forum National des DSI) and SN 2020 (Semaine du Numérique 2020). These events focused on artificial intelligence, big data and cyber-security. This paper is regarding to these topics.

References

[1] A. Perrot, *La visualisation d’information à l’ère du Big Data : résoudre les problèmes de scalabilité par l’abstraction multi-échelle*, thèse de doctorat, Université de Bordeaux. Bordeaux, France, 2017.

[2] J. Dean, S. Ghemawat, *Mapreduce: Simplified data processing on large clusters*, In 6th Symposium on Operating Systems Design and Implementation, USENIX Association, pages 137-150, 2004.

[3] Frédéric T.Ouédraogo, Abdoulaye Séré, Evariste Rouamba, Safiatou Soré, *Analysis of the 2015 Presidential Campaign of Burkina Faso Expressed on Facebook*,In 8th EAI International Conference on e-Infrastructure and e-Services for Developing Countries, AFRICOMM 2016. Ouagadougou, Burkina Faso, 2016

[4] Younghoon Kim, Kyuseok Shim, *Parallel top-k similarity join algorithms using MapReduce*, In IEEE 28th International Conference on Data Engineering, DOI:10.1109/ICDE.2012.87; 2012.

[5] Jeffrey Shafer, Scott Rixner, and Alan L. Cox. *The Hadoop Distributed Filesystem: Balancing Portability and Performance*, In 2010 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), DOI: 10.1109/ISPASS.2010.5452045, 2010.

[6] Moturi, Maiyo. *Use of MapReduce for Data Mining and Data Optimization on a Web*, International Journal of Computer and Applications (0975-8887) Volume 56– No.7, 2012.

[7] Apache : Apache Hadoop.http://hadoop.apache.org, 2010.

[8] Ouatik, Fahd and Erritali, Mohammed and Jourhmane, M., *Comparative study of MapReduce classification algorithms for students orientation*, Procedia Computer Science, page 1192-1197,volume 170, doi: 10.1016/j.procs.2020.03.030, 2020.

[9] Bagui, Sikha and Devulapalli, Keerthi and John, Sharon, *MapReduce Implementation of a Multinomial and Mixed Naive Bayes Classifier*, International Journal of Intelligent Information Technologies, pages 1-23, volume 16, doi : 10.4018/IJIIT.2020040101, 2020.

[10] N.Alamelu Menaka, Dr.Jabasheela, *Survey on Big Data processing using Hadoop, Map Reduce*, International Journal of Innovative Research in Information Security (IJIRIS) ISSN: 2349-7017(O) ISSN: 2349-7009(P) Volume 1 Issue 3, 2014.

[11] Abdoulaye SERE, Dario COLAZZO, Oumarou SIE, *A Hough Transform Based On a Map-Reduce Algorithm*,International. Journal of Engineering Research and Application ISSN : 2248-9622, Vol. 6, Issue 8 (part 2), 2016.

[12] Mateus Coelho ; Dylan Sugimoto ; Gabriel Melo ; Vitor Curtis and Juliana Bezerra *A MapReduce based Approach for Circle Detection*, In Proceedings of the 14th International Conference on Software Technologies - Volume 1: ICSOFT, 454-459, Prague, Czech Republic, 2019.

[13] Yaya TRAORE, Malo SADOUANOUAN, Didier BASSOLE, Abdoulaye SERE, *Multi-Label Classification using an Ontology*, International Journal of Advanced Computer Science and Applications (IJACSA), ISSN : 2156-5570 (Online) ISSN : 2158-107X (Print) Vol. 10, No. 12, DOI : 10.14569/issn.2156-5570, 2019.

[14] Apache : http://storm.apache.org/, 2020.

[15] Apache : https://flink.apache.org/, 2020.

[16] Apache : https://spark.apache.org/streaming/, 2020.

Predicting Hospitals Hygiene Rate during COVID-19 Pandemic

Abdulrahman M. Qahtani[1], Bader M. Alouffi[2], Hosam Alhakami[3], Samah Abuayeid[4], Abdullah Baz[5]
Department of Computer Sciences, College of Computers and Information Technology, Taif University
Taif, Saudi Arabia[1,2]
Department of Computer Science, College of Computer and Information Systems, Umm Al-Qura University
Makkah, Saudi Arabia[3,4]
Department of Computer Engineering, College of Computer and Information Systems, Umm Al-Qura University
Makkah, Saudi Arabia[5]

***Abstract*—COVID-19 pandemic has reached global attention with the increasing cases in the whole world. Increasing awareness for the hygiene procedures between the hospital's staff, and the society became the main concern of the World Health Organization (WHO). However, the situation of COVID-19 Pandemic has encouraged many researchers in different fields to investigate to support the efforts offered by the hospitals and their health practitioners. The main aim of this research is to predict the hospital's hygiene rate during COVID-19 using COVID-19 Nursing Home Dataset. We have proposed a feature extraction, and comparing the results estimating from K-means clustering algorithm, and three classification algorithms: random forest, decision tree, and Naive Bayes, for predicting the hospital's hygiene rate during COVID-19. However, the results show that classification algorithms have addressed better performance than K-means clustering, in which Naive Bayes considered the best algorithm for achieving the research goal with accuracy value equal to 98.1%. AS a result the research has discovered that the hospitals that offered weekly amounts of personal protective equipment (PPE) have passed the personal quality test, which lead to a decrease in the number of COVID-19 cases between the hospital's staff.**

Keywords*—*COVID-19; machine learning; hospitals hygiene; World Health Organization (WHO); personal protective equipment; K-means clustering; Naive Bayes; random forest

I. Introduction

Recently WHO has faced many challenges in increasing the global healthcare and Hygiene awareness to overcome COVID-19 pandemic. According to WHO COVID-19 is an infectious disease caused by a coronavirus, which started in December 2019 in the city of Wuhan in China [1]. On the other hand, hospitals and health practitioners are the most susceptible to infectious diseases. As a result, WHO has provided the Infection Prevention and Control (IPC) document [2], which consists of some required steps about water waste management, hand hygiene practices, safe healthcare waste management, safe management of dead bodies, and many other COVID-19 prevention requirements. Additionally, medical and health care fields are the most affected fields during COVID-19, which also consists of hospitals services and hospitals staff. However, there is a high demeaned in increasing the rate hospitals hygiene during COVID-19 pandemic to protect hospitals staff and hospital patients from COVID-19 infection. However, the availability of the personal protective equipment (PPE), such as hand sanitizer, Mask and gown, has become a worldwide concern, especially for health care workers. However, PPE can provide safe health-care services for hospital patients during COVID-19 pandemic. Moreover, WHO and all followed health ministries in whole world have increased there continues concern about offering all hygiene requirements, and tests for the hospitals to ensure of a high health safety level during the pandemic. However, the process of hospitals hygiene ensuring can take times and effort, but recently machine learning and deep learning has integrated in many COVID-19 researches to support health fields. relatively, introducing machine learning algorithms and methods in predicting the rate of hospitals hygiene can has great impact on enhancing hospitals and health fields services especially during COVID-19. The main contribution in this work is to build a machine learning model that can predict the rate of hygiene in hospitals during COVID-19, in which the most of the studies in the research area were focused on studying hand hand hygiene during COVID-19. As a sample we have used the COVID-19 Nursing Home Dataset, which consists of a USA counties hospitals monitoring information. The experiment has answered three major questions: which city has the most confirmed COVID-19 cases, which city has the most confirmed COVID-19 cases in the hospitals staff, which city has the high commitment in supply of PPE during COVID-19, is there any relation between COVID-19 cases and hygiene commitment. However, We have proposed a feature extraction, and comparing the results estimating from K-means clustering algorithm, and three classification algorithms: random forest, decision tree, and Naive Bayes, for predicting the hospital's hygiene rate during COVID-19 using COVID-19 Nursing Home Dataset. However, the results show that classification algorithms have addressed better in prediction than clustering, in which Naive Bayes considered the best algorithm for achieving the research goal. AS a result the research has discovered that the hospitals, that offered weekly amount of PPE have pass the personal quality test, which lead to decrease the number of COVID-19 cases between the hospitals staff. The work has structured as follows: Section 2 consists of the related studies, Section 3 includes the materials and methods. Section 4 is results and discussion. The last section concludes the work.

II. Related Studies

In most recent study [3] researcher conducted a review study for Coronavirus Disease-2019 (COVID-19), in his study he mentioned some recommendations, one of them is that people should be asked to practice hand hygiene frequently every 15–20 min. Hand hygiene compliance remains under the expectation and gets low among people who work in healthcare centres and visitors [4]. According to [5], statistics and recent studies indicate the rates of hand hygiene compliance in medical organisations, which hastate from 5 to 89%, with average compliance of 38.7%. The rate of Hand hygiene compliance almost stands on the method employed for measurement and place of assessment, with differentiation between clinics, intensive care units (ICU) and other places in the medical organisations. Several studies emphasis on that covering medical include its different unites by hand hygiene dispensers is not the ideal way to gain a high level of hygiene compliance [6][7]. That indicates the importance of using technologies to support decision-makers to allocate hand hygiene dispensers and get the vital data to monitor and manage hygiene systems in an optimal way [8]. [9] discuss in an explorer study several strategies and techniques for hand hygiene monitoring include direct observation, Video-assisted direct observation passive monitoring and automated individual monitoring. Furthermore, they reported and compared different hand hygiene monitoring systems in critical voice. In terms of using smart and automatic systems, several studies discuss and investigate using electronic solution for managing and monitoring hand hygiene compliance in hospitals and medical centers [10]. One of the applied solutions to improve the monitoring of hand hygiene system is an electronic device attached to a hand hygiene container to record the frequency of using that device. The device records each event by stamping time and date then transferred to a central computer to analysis and print a report related to that device. However, the electronic counting devices have a simple function to do it, which is counting events without more information like who has accessed the device, also cannot provide information about hand hygiene compliance [11]. The health sector is one of the main domains affected by those changes and takes advantages of that development in several medical fields, including sterilization within medical facilities. Much research conducted in this branch [12], [13]. For instance, [14] apply machine learning algorithm to investigate if location, time-based factors, or other behavior data can determine what characteristics are predictive of hand washing non-compliance events. Therefore, they found that the prediction of compliance can be done using those factors and would support decision-makers to make planes and decisions by using the predicted scenarios and results.

III. Materials and Methods

This section, have introduced a detailed explanation about the research methodology, dataset selection, the proposed model, data pre-processing, and programming tools. The work has been conducted on HP Pavilion laptop 8GB RAM, 1 TB HDD storage with windows 10 as an operating system. On the other hand, Anaconda with Jupiter Lab 1.2.6 has been used as the main platform for Python 3 programming language with sklearn library for K-means clustering and decision tree algorithms, and Rapid Miner 9.8 for random forest and naive Bayes algorithms. However, Rapid Miner is a data science and machine learning software that can deal with noisy data in a robotic manner. Moreover, open refine 3.4 has used for data cleaning, which explained in the next subsection as part of data pre-processing steps. The research has proposed a feature extraction and classification model for predicting the hospital's hygiene rate during COVID-19 using COVID-19 Nursing Home Dataset. However, the research has achieved by follow a well-defined methodology shown in Fig. 1, started by dataset selection, then data pre-processing, feature extraction, and attributes correlation study, after that one clustering algorithm: K-means clustering, and three classification methods: Decision Tree, Naive Bayes classifier, and Random Forest has applied, and performance has evaluated using the coefficient matrices.

A. Data

This subsection has presented information about the experiment dataset. The model has validated using the COVID-19 Nursing Home Dataset shown in Table I, which was offered by The Nursing Home COVID-19 Public File in, supported by The Centers for Medicare & Medicaid Services (CMS) and CDC's National Healthcare Safety Network (NHSN) system. The dataset consists of 87 columns with 245870 records that include different information about USA counties and provider cities, resident impact, facility capacity, staff, and supplies of protective equipment during the COVID-19 outbreak. Additionally, the first deadline for data reporting by USA counties hospital facilities was on the 17 of May 2020 at 11:59 p.m., where the dataset last updated date was the 17 of September 2020. However, data pre-processing and feature extraction have been explained in detail in the next section.

TABLE I. COVID-19 Nursing Home Dataset.

1- Dataset general information	
Total attributes	87
Total size	245870
Owner	CMS-Division of Nursing
Category	Special Programs-Initiatives
License	Public Domain U.S. Government
Contact Email	NH_COVID_Data@cms.hhs.gov
2- The experimental data	
Total attributes	12
Total size	120172
3- Extracted attributes description	
Pcity	Provider city
PQAT	Passed quality assurance test
WN95N	One week supply of N95 masks
WSM	One week supply of surgical masks
WEP	One week supply of eye protection
WGO	One week supply of gowns
WGL	One week supply of gloves
WHS	One week supply of hand sanitizer
TRC19	Total resident confirmed COVID-19
STC19	Staff total confirmed COVID-19
STD19	Staff total COVID-19 death
WTPPE	Weekly personal protection equipment

B. Data Pre-Processing and Feature Extraction

This section discussed the main steps of preparing the dataset before model implementation. Moreover, the feature extraction step, explained in this section is the main feature extraction, as a part of data processing, but in order to implement the models, the step has repeated in different ways depending on each model's requirements. Data pre-processing step has started by changing the dataset columns names in the Excel file, in which each column was referred as a

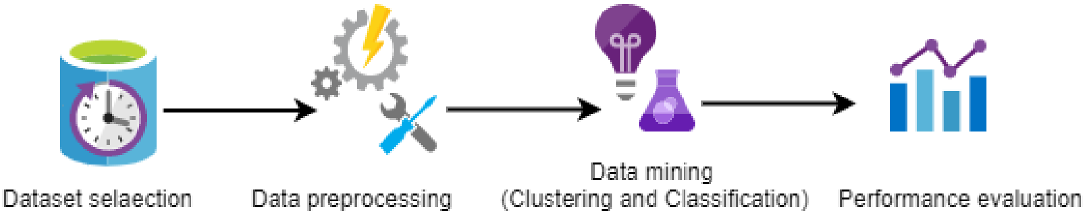


Fig. 1. Research Methodology

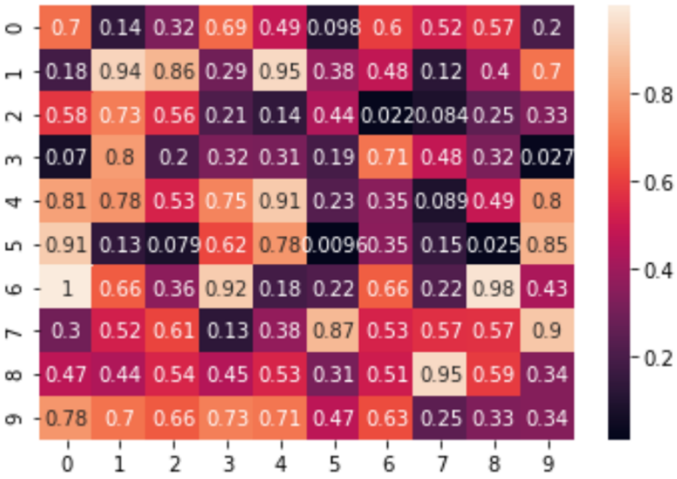


Fig. 2. The Relationship between Three Attributes: Staff Total Conferred COVID-19 Cases (STC19), Staff Total COVID-19 Death (STD19), and Total Resident Conferred COVID-19 Cases (TRC19)

combination of 3 to 4 words, the names has shorten by take the first letter in each word to simplify file reading step in the Python code Table I part 3 shows some examples of this step. The next step was eliminating the empty records and cells in the dataset, in which the COVID-19 Nursing Home Dataset was full of null values as a result from the empty records and cells. To illustrate this, open refine 3.4 has used to delete all the empty records and refill the empty cells with 'N' if the column data type is character, or '0' if the column data type is number. Additionally, during the model's implementation step, data transformation and replacement has done with the means of all other values, was the best solution. Furthermore, the outliers have eliminated from the dataset using Interquartile Range (IQR) in Python, and the outliers values have appeared in 13 columns with IQR values. Last step of data pre-processing was eliminating the duplicate values, which done on python, and decreased the total number of cleaned records to 120172 as shown in Table I. On the other hand, data reduction is one of the most important steps in any data pre-processing, in which it assists in reducing the data dimensionality, that lead to increasing model performance, identifying the irrelevant data, simplifying data visualisation, and enhancing results predicting. However, feature extraction is one of the well-known methods in data reduction, and we have applied two techniques of feature extraction depending on the dataset needs. We have started by feature selection, or it can refers as attributes selection as that the research did not follow any feature selection approaches or algorithms the step has done by selecting the target attributes from the dataset depending on dataset knowledge and the experimental requirements. The extraction was achieved using a python script in JupyterLab, and we have successfully extracted 11 attributes from the dataset as shown in Table I (3). Furthermore, feature construction has done in two steps on a specifies dataset attributes: WN95N, WSM, WEP, WGO, WGL, WHS, Firstly, we have transformed the values in each attribute to '0' for 'N', and '1' for 'Y' so it became a numerical values, then we have applied an aggregation process on the attributes, because they all representing the availability of PPE, which represented in one attribute: weekly total personal protective equipment (WTPPE), that has values in the interval of 0 to 6, shown in Table I (3). Before implementing the model a 500 objects has extracted from the dataset as a sample for studying the correlation and relationship among dataset attributes. However, multiple types of visualization have used in JupyterLab such as horizontal line charts and heatmap. To illustrate this, the horizontal line chart in Fig. 3 shows the relation between USA counties provider cities, counties hospital staff total conferred COVID-19 cases in the blue line, and counties residents total conferred COVID-19 cases in the red line. The curves implied a high addressing rate in staff cases, and residents in Shawnee counties, and a stabilized rate from Appleton to Drumright, where the curve returned to increase in Portland for staff cases. Consequently, the curves in Fig. 3 have come to an agreement, in that Shawnee counties have addressed the highest COVID-19 cases whether in hospitals staff, or counties residents, also the curves show the same stabilized rate in some cities, which implied a medium to high correlation between these attributes. Depending on the results presented in Fig. 3, a 10*10 Heatmap shown in Fig. 2 indicating a high correlation between three attributes: staff total conferred COVID-19 cases (STC19), staff total COVID-19 death (STD19), and total resident conferred COVID-19 cases (TRC19) confirmed. However, data description and correlation study step has been accomplished as part of data pre-processing, because each one of these attributes can give a specified descriptive for COVID-19 situations inside and outside the hospitals, which can be linked with the level of hospital hygiene.

C. K-Means Clustering

This section introduced the K-means clustering model as a unsupervised learning algorithm, which was implemented in our research. However, among all data mining research, clustering analysis research is the most influential one because naturally things and people are clustered in multiple classes

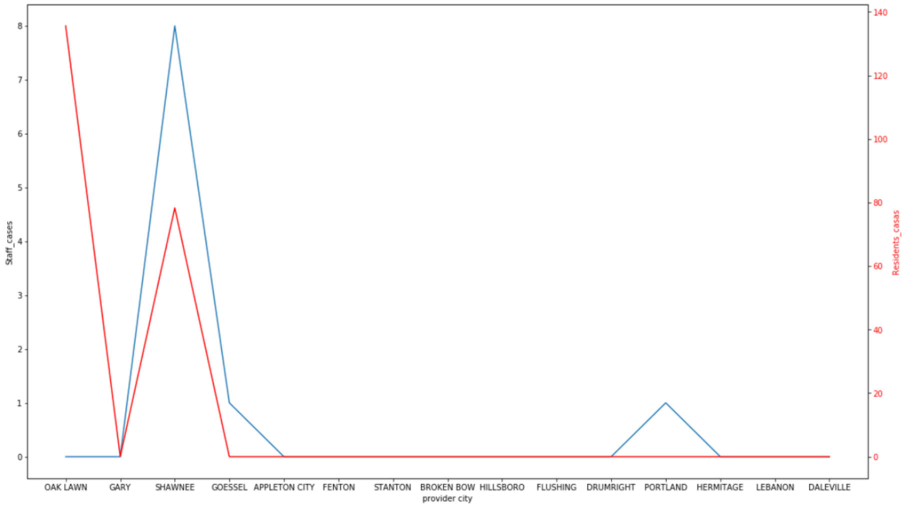


Fig. 3. The Chart Shows Three Different Attributes Provider Cities Names in the X axis, Hospital Staff Total Conferred COVID-19 Cases in the Blue Line, and Countries Residents Conferred COVID-19 Cases in the Red Line

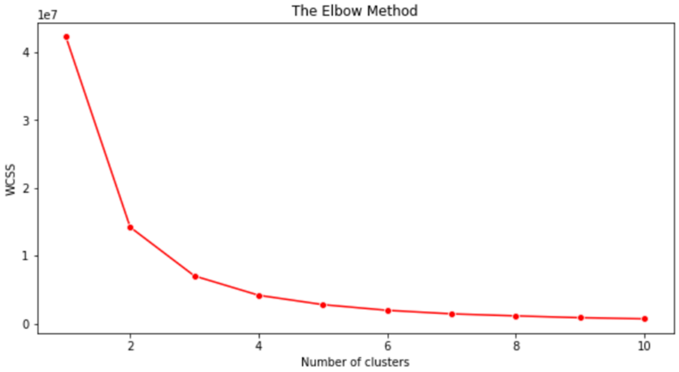


Fig. 4. The K-WCSS Curve Shows an explicit Inflection when K = 2 for the Elbow Method

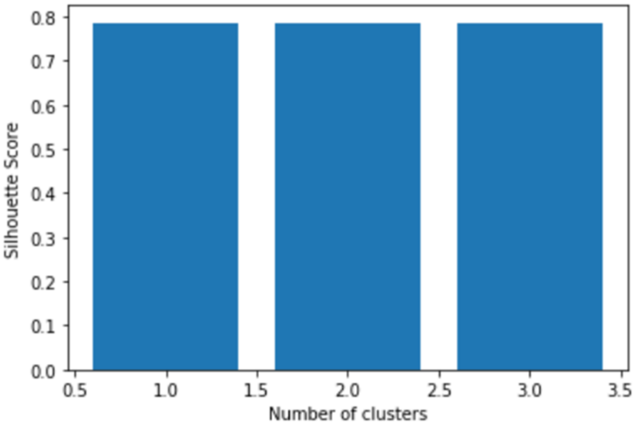


Fig. 5. The Silhouette Score Method Results (When K=1 and K=2)

and groups. Relatively, K-means clustering is one of the most used clustering methods for its fast convergence and simplicity. However, it's an iterative algorithm that uses the distance as a metric, classes of K in the data set, which represent the number of clusters, the mean of the distances, and a giving initial centroid for each class. The main problem in K-means implementation is selecting the appropriate K-value, which directly affects the algorithm result. The study [15] argued on the types of K-value selecting algorithms, such as that the Elbow , silhouette coefficient, gap statistic, and canopy. The author found that Elbow Method records the least execution time among the other methods. In this work a K-means clustering has built to cluster the hospital's hygiene depending on two attributes: PQAT, and WTPPE, that shown in Table I (3). In the first step the most relevant dataset attributes for our target attribute PQAT has examine, and that has been achieved by features selection methods. However, the ”SelectKBest” method from ”sklearn” library in Python has used to discovered 4 features the best relevant to PQAT: PCity, STC19, STD19, WTPPE, and WTPPE has chosen to be the second attribute in the 2D K-means model, in which the total size of records used in K-means clustering was 70673. Then as inspired from [15], we performed a comparison between two of K-values estimated methods: the elbow method, and silhouette score method. The main idea behind the elbow method is using, the square of the distance between sample points. Each cluster and its centroid gives a series of K values. The sum of squares error (SSE), or the within cluster sum of squares (WCSS) are used as a performance indicator, and Iterate the value of K to calculate the SSE or WCSS, where the smaller value of SSE or WCSS represent, that each cluster is similar. The K-WCSS curve in Fig. 4 shows the result of applying the elbow method on our model, and the curve indicated an explicit inflection in K = 2. On the other hand, the silhouette score method measures the similarity of an object to its cluster compared to the other clusters, where silhouette score high value indicates that a well matched object in the same clusters, and poorly matched with the other cluster. Fig. 5, shows the result of calculating the best number of k-values that can enhance the performance of our K-means clustering model.

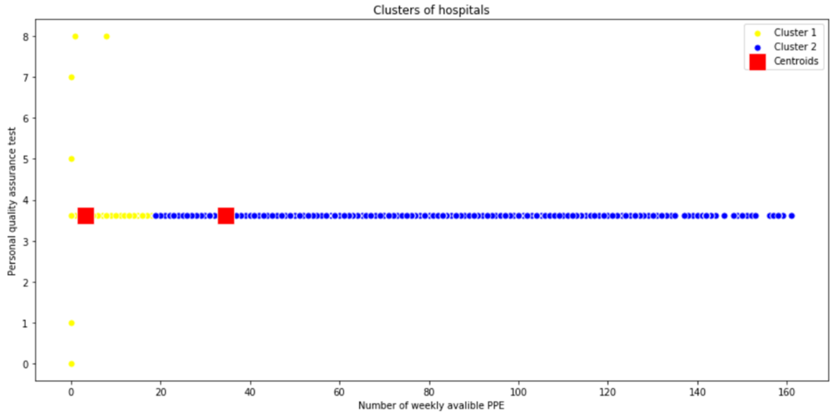


Fig. 6. The Hospitals Clusters according to Weakly available PPE, and Personal Quality Assurance Test

D. K-Means Clustering Measurement Metrics

This section explained the measurement metrics used to evaluate the K-means clustering model. This work has used six different clustering K-means clustering:

- **Adjusted rand index (ARI):** Measured the corrected version of rand index, which is a statistical measurement that represents the similarity between two data clustering. The ARI value establishes a minimum or starting point for comparisons between two pairs of clusters.
- **Mutual Information based scores (MI):** Statistically it measures the strength of association between two variables.
- **Homogeneity, completeness and v-measure (HCV):** Represent the quality of clustering algorithms.
- **Fowlkes-mallows scores (FMS):** Evaluate the similarity between clusters.
- **Calinski-harabaz index (CHI):** It is the ratio of the sum among the clusters, the value indicates better algorithm performance.
- **Davies-bouldin score (BD):** represents the ratio among cluster scatter, and cluster separation, often used to evaluate the best number of clusters, lower value indicates better clustering.

The K-means algorithm has repeated in the range of K-values from 1 to 2, and each time the algorithm measurement metrics have evaluated using the 6 previous metrics, which are shown in Table II (2). According to Table II, using 2 clusters is more suitable with our data. However, according to Fig. 6 there were two clusters for the hospitals PQAT, and WTPPE, but because of how the clusters have visualized, in which each cluster appear in one line, which emphasized that the K-means clustering is not suitable on the attributes (PQAT, and WTPPE).

E. Classification Methods

This section have introduced in detail the classification methods that have been implemented in this work. Classification is one of data mining techniques that aim to determine dataset records classes depending on the value of the predefined target attribute from the dataset. However, there are many classification methods that proved their benefits in prediction and identified various life problems. To illustrate this, Ira Ekanda Putri [16] used two different classification methods: naive Bayes classifier, support vector machine (SVM) to predicate heart disease, in which SVM was addressed the best prediction performance. On the other hand, Dedy Hartama [17] used the C4.5 decision tree to predicate patterns of interest of high school graduates. The study was conducted on a new admission student's dataset consisting of: student ID, student name, school location, and school type. The experiment was done on a rapid miner studio to build the C4.5 decision tree, which records a good performance in classifying students into their interesting study program according to the given attributes in the dataset. Moreover, unlike the supervised data mining methods such as clustering algorithms, the classification methods are supervised methods, which need to divide the dataset into train and test, and use labeled classes for model training steps. In this work three classification methods have chosen: Decision Tree, Naive Bayes classifier, and Random Forest, and the dataset has divided into 70% train data and 30% for the test data.

TABLE II. THE EXECUTION TIME FOR K-VALUE ESTIMATING METHODS AND K-MEANS CLUSTERING MEASUREMENT METRICS.

1-Methods execution time		
K-value	Elbow method	Silhouette score
1	1.6 s	-
2	2.4 s	1388.2 s
2- Measurement metrics		
K-value	Metrics	Score
1-2	ARI	0 - 0.29
1-2	MI	-1.2e-15 - 0.43
1-2	HCV	1 - 1
1-2	FM	0.6 - 0.68
1-2	CH	1.59e+06 - 1.59e+06
1-2	DB	0.5 - 0.5

*1) **Decision Tree**:* The first classification model used in this work is the decision tree, which is one of the most known classification methods. However, decision trees have

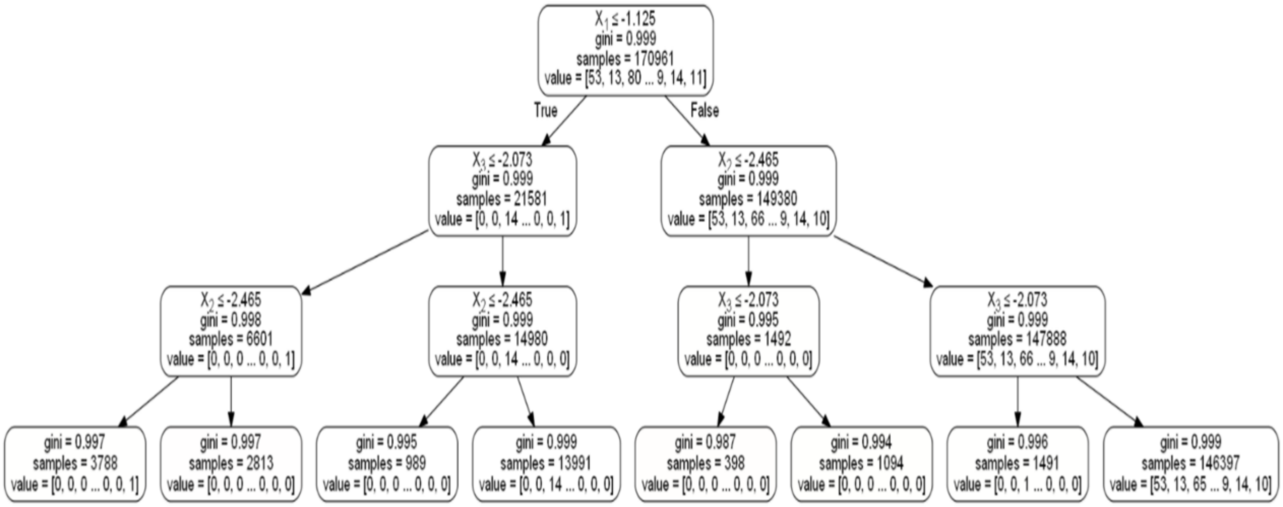


Fig. 7. Decision Tree

many advantages compared to other classification methods, in which it is inexpensive to construct, extremely fast at classification of unknown records, and easy to interpret for small dataset, but it can take time in the training step. In this work the decision tree algorithm has applied on COVID-19 Nursing Home dataset using Jupyter Lab 1.2.6 as a python 3 programming environment. Additionally, four attributes have extracted from COVID-19 Nursing Home Dataset, which are SQT, STD19, TRD19, and STC19. The algorithm starts by selecting the best split and when to stop splitting for the tree. One of the most known measurements used to measure the node impurity, and the best split of the decision tree is Gini index (Gini coefficient). Additionally, the Gini index shown in Equation 1, is a measurement, which calculates the probability rate of a particular feature, that is wrongly classified when randomly selected.

$$Giniindex = 1 - \sum_{i=1}^{n}(p_i)^2 \qquad (1)$$

However, Fig. 7 shows the resulting decision tree that was built to predict the target attribute PQT according to STD19, TRD19, and STC19. Additionally, the Gini index is estimated for each node in the tree, and in each branch the tree takes a sample from the dataset. However, the branches have been labeled into true or false. However, the tree has stopped splitting with 8 leafs which presented the final prediction results.

2) ***Naive Bayes classifier:*** The second classification method, which has been used in this work is Naive Bayes classifier. Naive Bayes classifier is a straightforward classification method that does not require any complex rules for implementation. To illustrate this, Naive Bayes depend on probability theory for finding the best classification class for the target attribute. However, the main advantages of Naive Bayes, that it is a robust method for isolating noise points in the dataset, handling the missing values by ignoring the missing values during calculating the probability, and it has the ability for dealing with the irrelevant attributes. A study by Aji Prasetya Wibawa [18] has proposed a Naive Bayes model for classifying journals Quartile. The study included 1491 records and 10 attributes collected from "Journal Rankings in the Scimago Journal and Country Rank", and specified in computer science. The model was addressed with a good prediction rate with accuracy approximately equal to 71.60%, which concluded that Naive Bayes had the ability to classify journals Quartile even if that the accuracy was not optimal. On the other hand, in this work Naive Bayes has applied on the dataset using Rapid Miner 9.8, and three attributes have selected: PQT, WTPPE, and STC19, in which PQT is the target and WTPPE, STC19 are the predictor attributes. Fig. 9 shows the results of Naive Bayes classifier in predicting the PQT using WTPPE and STC19. To illustrate this, in Fig. 9(A) the result shows that the hospitals that have high availability of personal protective equipment equal to 6 have high probability to pass the personal assurance test, where hospitals that have low than 6 WTTP have high probability to fail in the PQT. On the other hand, Fig. 9(B) shows that there is increasing rate in the probability of pass the PQT when the number of conferred COVID-19 cases in the hospitals staff (STC19) is less than 25, then the probability of pass decreased while the number of STC19 increased. Additionally, in Fig. 8 we have examined

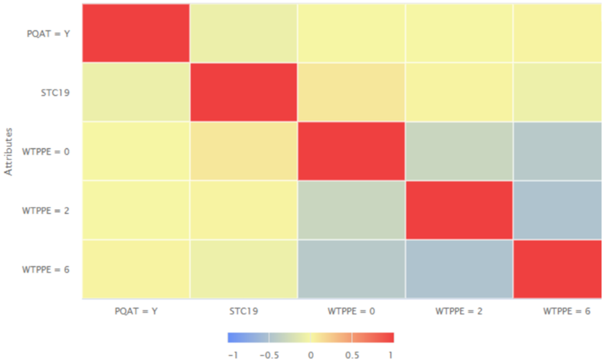


Fig. 8. Naive Bayes Classifier Correlation Matrix for Three Attributes: PQT, WTPPE, and STC19

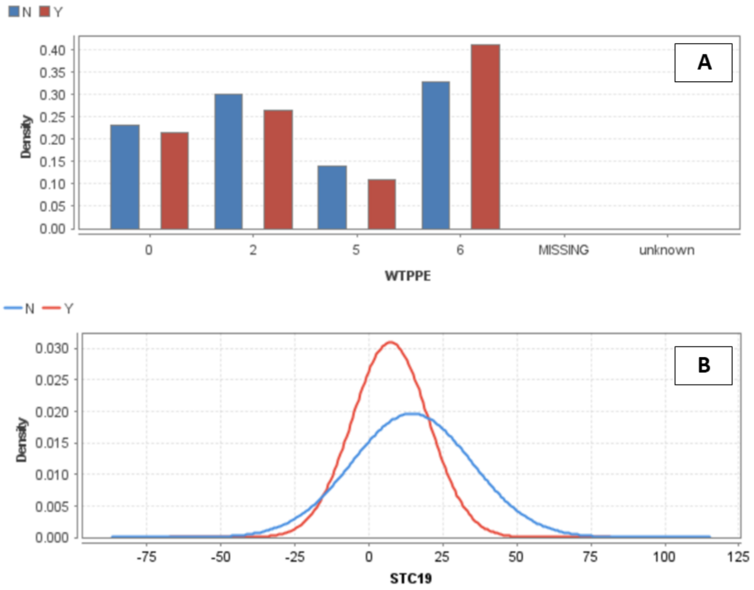


Fig. 9. Naive Bayes Classifier Result, (A) Predicting PQT Attribute depend on WTPPE Attribute, (B) Predicting PQT Attribute depend on STC19 Attribute

the correlation between the three attributes used in the Naive Bayes classifier, the heatmap shows a high correlation between PQT, WTTP, and STC19.

3) ***Random forest :*** The last classification model, that has applied on the dataset is random forest classifier. However, random forest is a combination of predictors trees, in which the value of each tree depends on a random vector sampled separately and with the same distribution for all trees in the random forest. The algorithm was discovered by Leo Breiman [19]. Moreover, a study by Ramón Díaz-Uriarte [20], has used a random forest algorithm to select the relevant gene expression for sample classification, such as distinguish between cancer and non-cancer patients. The author argued that random forest recorded an optimal performance in predicting and classification genes expression, and it has the ability to deal with noisy data such as the micro-array data. For this work the random forest model have applied for predicting the target attribute PQT using the STC19 attribute using Rapid Miner 9.8. However, Fig. 10 shows a snapshots of the resulted random forest, in which 20 trees have applied in the predicting model, each tree has classified the hospital's hygiene according to PQT. The figures show that the trees from 1 to 12 have predicted that most hospitals with staff confirmed COVID-19 cases have passed the personal quality test, except if the staff cases are larger from 137,500 or less than 102.500 then the hospitals have not passed the PQT. The result obtained in tree 12 in Fig. 10 has the same result obtained previously by the decision tree model in Fig. 7. However, at tree 20 in Fig. 10 all the hospitals have passed the PQT, in which all the tree labels (leafs) became equal 'Y', and that could be not a reasonable decision.

F. Classification Models Evaluation

This section have explained in detail the performance evaluation step for the classification models used in this research. The evaluation was done using four performance metrics: accuracy, F1 score, recall, and precision.

TABLE III. The values of classification models evaluation metrics

Model	Accuracy	F1 score	Recall	Precision
Decision tree	0.1	2.1	1.2	0
Naive Bayes	98.1	99	100	98.1
Random forest	98.1	0	100	98.1

- **Accuracy:** It is defined by the ratio of the correct predicted observation of the model to the total of observations.
- **Recall:** Defined as the ratio of the correct predicted of the positive observations to the total observations in the actual class.
- **Precision:** Defined as the ratio of the correct predicted of the positive observations to the total predicted positive observations.
- **F1-Score:** It presented the average of Precision, Recall and described by Equation 2

$$2 * (Recall * Precision)/(Recall + Precision) \quad (2)$$

However, Table III shows the results of the evaluation metrics on the three classification models: decision tree, Naive Bayes, and random forest. The results shows that Naive Bayes and random forest have addressed a high accuracy in prediction of hospital hygiene with values equal to 98.1 compared to decision trees. Moreover, Naive Bayes and random forest have recorded the same value for recall and precision metrics, which equals to 100 and 98.1 respectively, where decision tree has addressed the lowest values for recall and precision metrics equals to 1.2 and 0 respectively. However, according to F1 score values we have realized that Naive Bayes classifier is the convenience classification algorithm for prediction hospitals hygiene, in which that Naive Bayes has record the highest f1

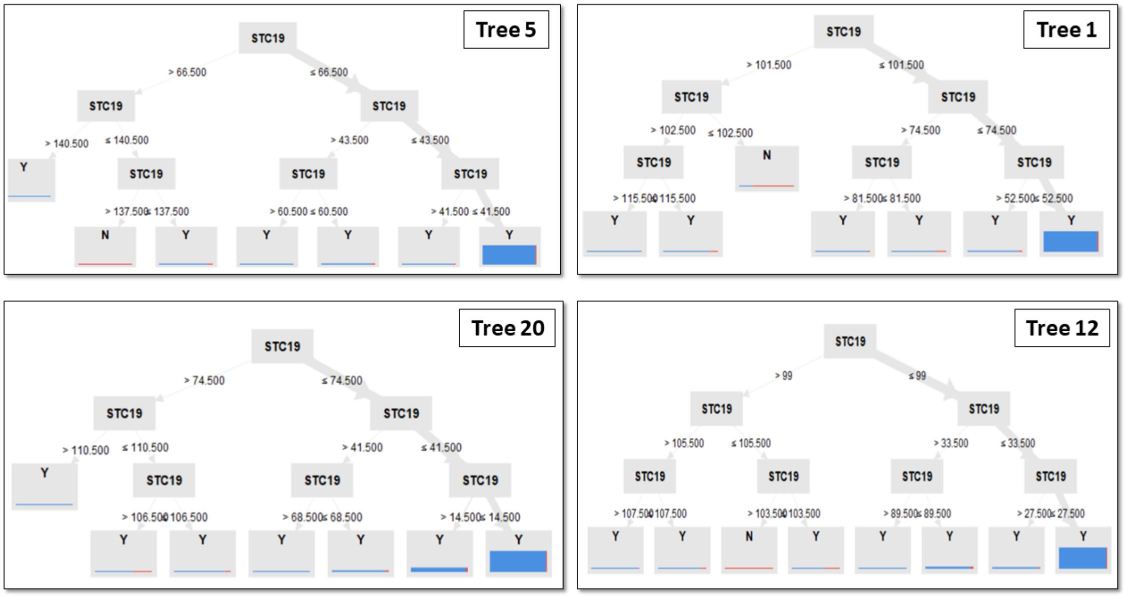


Fig. 10. Sample from Random Forest Results

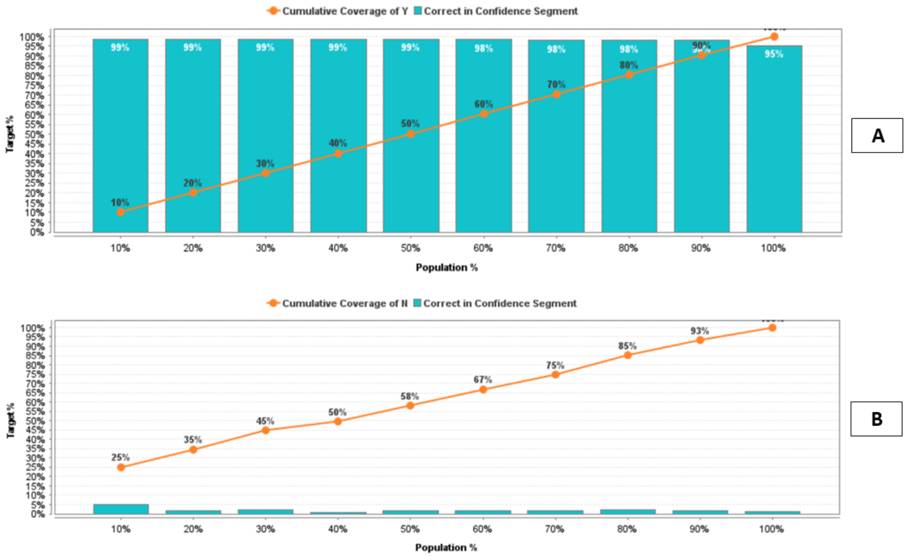


Fig. 11. Life Chart (A) Naive Bayes Classifier, (B) Random Forest

score equals to 99, where the other classification algorithms have addressed 2.1 for decision tree and 0 for random forest. On the other hand, Fig. 11 shows the life chart for both Naive Bayes and random forest, in which Naive Bayes life chart has recorded the best prediction and coverage values. To illustrate this, in Fig. 11 (A) the ratio of correct confidence and prediction with coverage have reached to the highest values with the increase in the life chart, in which the coverage has reached to 100%, and the correct confidence has reached to 95%. However, Fig. 11 (B) shows a random forest low value for the correct confidence equals to %5 corresponding to a high value for coverage equals to %100. As a result according to Table III and Fig. 11 we have considered, that Naive Bayes is the best classifier in prediction hospitals hygiene on the COVID-19 Nursing Home dataset.

IV. Results and Discussion

This section have discussed the results obtained from K-means clustering as a clustering algorithm, and Naive Bayes as classification algorithms, which has recorded the best prediction result compared to random forest and decision trees. The aim from this research is to predict the hospital's hygiene rate during COVID-19 using COVID-19 Nursing Home Dataset. According to the result from K-means clustering on hospitals PQAT, and WTPPE, which shown in Fig. 6, the research has emphasized that the K-means clustering is not suitable for the research problem. Additionally, the main limitation of K-means clustering is the long execution time required to implement the algorithm on a big dataset. On the other hand, according to Table III and Fig. 11, Naive Bayes is the best classifier in prediction hospital hygiene on the COVID19 Nursing Home dataset. As a result Naive Bayes is the best algorithm that is suitable to the research dataset, and to study the hospital's hygiene. The research has discovered that the hospitals that offered weekly amounts of PPE have passed the personal quality test, which lead to a decrease the number of COVID-19 cases between the hospital's staff.

V. Conclusion

This research, have proposed a feature extraction, and comparing the results estimating from K-means clustering algorithm, and three classification algorithms: random forest, decision tree, and naive Bayes, for predicting the hospital's hygiene rate during COVID-19 using COVID-19 Nursing Home Dataset. However, most of the studies in the research area were focused on studying hand hand hygiene during COVID-19. Additionally, the results show that classification algorithms have addressed better in prediction than clustering, in which naive Bayes considered the best algorithm for achieving the research goal with accuracy value equal to 98.1%. As a conclusion from this research hospitals that offered weekly amounts of PPE have passed the personal quality test, which lead to a decrease in the number of COVID-19 cases between the hospital's staff.

VI. Acknowledgment

The authors extend their appreciation to The Researchers Supporting Project number (1-441-68) Taif University, Taif, Saudi Arabia

References

[1] S. Chen, J. Yang, W. Yang, C. Wang, and T. Bärnighausen, "Covid-19 control in china during mass population movements at new year," *The Lancet*, vol. 395, no. 10226, pp. 764–766, 2020.

[2] W. H. Organization *et al.*, "Water, sanitation, hygiene, and waste management for sars-cov-2, the virus that causes covid-19: interim guidance, 29 july 2020," World Health Organization, Tech. Rep., 2020.

[3] Z. Y. Zu, M. D. Jiang, P. P. Xu, W. Chen, Q. Q. Ni, G. M. Lu, and L. J. Zhang, "Coronavirus disease 2019 (covid-19): a perspective from china," *Radiology*, p. 200490, 2020.

[4] S. Debnath, D. P. Barnaby, K. Coppa, A. Makhnevich, E. J. Kim, S. Chatterjee, V. Tóth, T. J. Levy, M. d Paradis, S. L. Cohen *et al.*, "Machine learning to assist clinical decision-making during the covid-19 pandemic," *Bioelectronic medicine*, vol. 6, no. 1, pp. 1–8, 2020.

[5] J. Kirk, A. Kendall, J. F. Marx, T. Pincock, E. Young, J. M. Hughes, and T. Landers, "Point of care hand hygiene—where's the rub? a survey of us and canadian health care workers' knowledge, attitudes, and practices," *American journal of infection control*, vol. 44, no. 10, pp. 1095–1101, 2016.

[6] P. Parchure, H. Joshi, K. Dharmarajan, R. Freeman, D. L. Reich, M. Mazumdar, P. Timsina, and A. Kia, "Development and validation of a machine learning-based prediction model for near-term in-hospital mortality among patients with covid-19," *BMJ Supportive & Palliative Care*, 2020.

[7] H. Burdick, C. Lam, S. Mataraso, A. Siefkas, G. Braden, R. P. Dellinger, A. McCoy, J.-L. Vincent, A. Green-Saxena, G. Barnes *et al.*, "Prediction of respiratory decompensation in covid-19 patients using machine learning: The ready trial," *Computers in biology and medicine*, vol. 124, p. 103949, 2020.

[8] L. J. Conway, "Challenges in implementing electronic hand hygiene monitoring systems," *American Journal of Infection Control*, vol. 44, no. 5, pp. e7–e12, 2016.

[9] A. Marra and M. Edmond, "New technologies to monitor healthcare worker hand hygiene," *Clinical Microbiology and Infection*, vol. 20, no. 1, pp. 29–33, 2014.

[10] J. M. Boyce, "Current issues in hand hygiene," *American journal of infection control*, vol. 47, pp. A46–A52, 2019.

[11] S. Hagel, J. Reischke, M. Kesselmeier, J. Winning, P. Gastmeier, F. M. Brunkhorst, A. Scherag, and M. W. Pletz, "Quantifying the hawthorne effect in hand hygiene compliance through comparing direct observation with automated hand hygiene monitoring," *infection control & hospital epidemiology*, vol. 36, no. 8, pp. 957–962, 2015.

[12] J. Conway, "The industrial internet of things: an evolution to a smart manufacturing enterprise," *Schneider Electric*, 2016.

[13] V. Tsiatsis, S. Karnouskos, J. Holler, D. Boyle, and C. Mulligan, *Internet of Things: technologies and applications for a new age of intelligence*. Academic Press, 2018.

[14] P. Zhang, J. White, D. Schmidt, and T. Dennis, "Applying machine learning methods to predict hand hygiene compliance characteristics," in *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2017, pp. 353–356.

[15] C. Yuan and H. Yang, "Research on k-value selection method of k-means clustering algorithm," *J—Multidisciplinary Scientific Journal*, vol. 2, no. 2, pp. 226–235, 2019.

[16] I. E. Putri, D. Rahmawati, and Y. Azhar, "Comparison of data mining classification methods to detect heart disease," *Pilar Nusa Mandiri: Journal of Computing and Information System*, vol. 16, no. 2, pp. 213–218, 2020.

[17] D. Hartama, A. P. Windarto, and A. Wanto, "The application of data mining in determining patterns of interest of high school graduates," in *Journal of Physics: Conference Series*, vol. 1339, no. 1. IOP Publishing, 2019, p. 012042.

[18] A. P. Wibawa, A. C. Kurniawan, R. P. Adiperkasa, S. M. Putra, S. A. Kurniawan, Y. R. Nugraha *et al.*, "Naïve bayes classifier for journal quartile classification," *International Journal of Recent Contributions from Engineering, Science & IT (iJES)*, vol. 7, no. 2, pp. 91–99, 2019.

[19] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[20] R. Díaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, vol. 7, no. 1, p. 3, 2006.

A Comparison of EDM Tools and Techniques

Eman Alshehri[1], Hosam Alhakami[2], Abdullah Baz[3], Tahani Alsubait[4]
College of Computer and Information Systems
Umm Al-Qura University
Makkah, Saudi Arabia

***Abstract*—Several higher educational institutions are adapting the strategy of predicting the student's performance throughout the academic years. Such a practice ensures not only better academic outcomes but also helps the institutions to reorient their curriculums and teaching pedagogies so as to add to the students' learning curve. Educational Data Mining (EDM) has risen as a useful technology in this league. EDM techniques are now being used for predicting the enrolment of students in a specific course, detection of any irregular grades, prediction about students' performance, analyzing and visualizing of data, and providing feedback for overall improvement in the academic spheres. This paper reviews the studies related to EDM, including the approaches, data sets, tools, and techniques that have been used in those studies, and points out the most efficient techniques. This review paper uses true prediction accuracy as a standard for the comparison of different techniques for each EDM applications of the surveyed literature. The results show that the J48 and K-means are the most effective techniques for predicting the students' performance. Furthermore, the results also cite that Bayesian and Decision Tree Classifiers are the most widely used techniques for predicting the students' performance. In addition, this paper highlights that the most widely used tool was WEKA, with approximately 75% frequency. The present study's empirical assessments would be a significant contribution in the domain of EDM. The comparison of different tools and techniques presented in this study are corroborative and conclusive, thus the results will prove to be an effective reference point for the practitioners in this field. As a much needed technological asset in the present day educational context, the study also suggests that additional surveys are recommended to be driven for each of the EDM applications by taking into account more standards to set the best techniques more accurately.**

Keywords*—Educational Data Mining (EDM); students' performance; prediction; higher education; WEKA

I. Introduction

Data mining is the most effective process for analyzing big data warehouses to derive valid and useful information, to extract hidden data, and to detect relationships between factors in massive data [1]. The educational data mining (EDM) process uses computational methods to convert raw data from educational systems into useful information to help educational issues [2]. Education nowadays contains several enhancement methods used to supervise and identify the students' academic performance in their studies. Data mining has been considered as one of the most useful processes used to identify students' performance. Presently, the scope of Data Mining has not limited to education only as it covers almost all those domains where data is used. There are many examples of applications using data mining. Retail management is one of such applications. Other examples include applications within banking sector, telecommunication, marketing, hospitality, production management, and so on. These organizations take the benefits from data mining to increase their income and future growth [3]. The data extraction in the field of education through using data mining methods is typically known as Educational Data Mining (EDM). Nowadays, EDM is a new discipline concerned with a different approach [1][4]. In the current context, education provides various ways and systems of learning that students can access. To quote a few, these include: Learning Management System (LMS) which is so popular and needed these days along with conventional classroom learning and Learning Object (LO). Social networking and online forums are other also needed in the E-learning process. Adaptive Hypermedia systems, educational games, concept maps and online exams are other points of educational contact needed by students worldwide. Each of these platforms brings several types of data, which EDM has to handle. [5]. Many educational institutions assess the students' performance depending on the course content and knowing the objectives to fulfill an effective learning process [4]. One of the biggest goals of higher education institutions is to improve and enhance the quality education process for its students. One way to improve the quality level in the education system is by discovering and applying data mining techniques [6]. Data mining is used to predict students' registration in a certain course and to detect any abnormal values of grades, prediction about students' performance, analysis and visualization of data, providing feedback to support instructors, recommendations for students, and so on [7][8]. Data mining techniques also assist in advising students in choosing the appropriate subjects for their undergraduate or postgraduate courses in the university. Data mining discovery has become an area of growing importance, especially in education, as it assists in students' data analysis by using several factors and interpreting it to deliver a useful information [3].

This paper surveys literature review regarding EDM with data sets size and techniques used in such studies. The present study also aims at identifying the most effective technique for Educational Data Mining. This paper is divided into five sections. Section 1 shows an introduction for the paper purpose and structure. Section 2 examines the background of Educational Data mining (EDM) methods. Details about phases of data mining are shown within Section 3. In Section 4, we have discussed some applications of educational data mining, while Section 5 discusses the results. Section 6 concludes this research and posits suggestions for future work in the same domain.

II. Background

EDM certainly helps in reaching the needed goals for the educational process. By applying EDM methods, we can build prediction models for enhancing students' performance [9].

A. Data Mining Techniques

There are many research papers and studies regarding the use of data mining techniques in education. The most common and widely used techniques for predicting students' performance are regression and classification, but other techniques have also been used, such as Clustering [10]. The EDM is useful for improving the process of studying, advising students, finding the reasons leading to dropouts, predicting students' performance, detection of undesirable behaviour. EDM can also help the educators to track academic progress to improve the teaching process. These algorithms in data mining require a quick mention to be familiar with [11]. A list of techniques explanation is stated below:

- **Classification** is one of the data mining applications that divides data into target classes [12]. The classifier algorithm uses a pre-classified prototype for identifying the set of parameters required in classification, for allocating a category to a record. The classification aims to predict the target class for each status of data accurately [13]. In EDM, this technique is used for classifying students based on the characteristics such as age, gender, grades, behavior, etc. The major classification algorithms are BayesNet, Naïve Bayes, C4.5 (J48), ID3 and Neural Network (NN).
    - **J48** classifier is a kind of decision trees algorithms. It consists of many nodes starting from the root till ending with the leaf. This algorithm can fix the issue of overfitting data and un- pruning. It is also able to specify the attributes are relevant or irrelevant at classification. In each node of the tree, J48 chooses the best effective feature to divide its sample into subsets at different classes. J48 can also deal with continuous and discrete data. [14]. Also, J48 algorithm repeatedly classifies data until it reaches the optimum level of categorization.
    - **Naive Bayes** classification is a simple classification algorithm that can calculate the probability by calculating the combination of entries and frequency in the data set [15].
- **Regression** technique is mainly used when there is a need to predict how one or more independent variables are related to dependent ones. Dependent variables are the ones to be predicted, while independent variables are known in prior [13] [16]. In EDM, it is used for the prediction of students' academic performance, prediction of the final grade, etc. Support Vector Machine (SVM), linear regression and neural networks are some common methods for regression within educational data mining [17]. Moreover, Classification and Regression Trees can be used together at the same model like CART technique.
- **Clustering** splits the data set into different groups, known as clusters. Clustering is needed mainly in cases where the ultimate usual data set categories within the data set are not known previously. The data point within a cluster should be same as other data points within the same cluster and different from data points of different cluster [18] [19]. In EDM, the clustering technique is used for grouping according to similarities and differences between students, courses, behavior, etc. [13]. The most popular clustering methods are K-mean and X-mean [20].
- **Decision Trees** are used for applying the classification model in the form of a tree structure. Each inner tree node represents examination for attribute, while a branch is a symbol for the result of that examination, and each leaf node acts as a classification. The classification rule is a path from the root to the leaf. Stability and easy interpretation is the biggest advantage of this technique. It is also suitable for solving different problems in various sectors, such as finance, business, education, etc.

B. Data Mining Tools

There are several tools that help in data mining and most of them are open source. An exhaustive perusal of the literature in this context helped us to list out the tools that have been used frequently in different research studies. The tools are listed below:

- **WEKA** is a java based tool used to process big data sets. It includes different algorithms, which may be applicable within data mining techniques [21]. It can be easily applied to algorithms to obtain quick results [22].
- **RapidMiner** is a tool developed by the rapid miner company. It provides features for machine learning, data mining, etc. It is mainly used for research, rapid prototyping, training, and of course education. It assists all the phases of data mining, including results such as validation, optimization and visualization [23].

III. Data Mining Process

The data mining process consists of phases that allow us to convert unknown information (raw data) into valid and meaningful information (knowledge) [24]. The following Fig. 1 shows knowledge extraction in sequential steps that are part of data mining process.

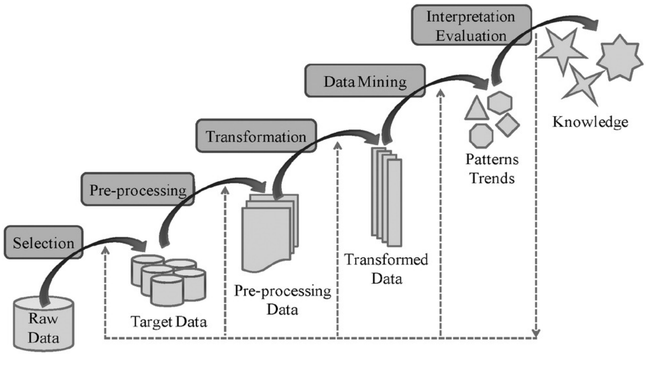


Fig. 1. Data Mining Process[25].

1) Data Selection: This step involves selecting or retrieving a data set in which the process of discovery has to be performed as a pre-processed data [26].
2) Data Pre-Processing: This step involves making the data more reliable in this stage, i.e., remove irrelevant data from the data set, and find the missing values and handle it [27].
3) Data Transformation: In this step the data is transformed and is categorised into appropriate formats for mining, i.e., performing some algorithms as classification and clustering [27].
4) Data Mining: It is one of the most significant steps of the process in which techniques and tools are applied to extract useful patterns. Data mining algorithms include classification, clustering, regression, etc. [28].
5) Pattern Evaluation: In this stage, we can identify specific patterns and evaluate to come up with the desired goals.
6) Knowledge Representation: This is the last phase where the knowledge obtained previously is visually represented to the user. This stage makes use of visualization techniques to help the users to have full images of the results and interpret the outcomes [26].

IV. Literature Review

The literature review undertaken for the present research context specifically includes research studies published in the last five years. These studies have assisted in predicting students' performance through EDM techniques. Moreover, the analysis also revealed different educational data mining techniques that are being used, tools, data set size, the best algorithm used with the highest prediction accuracy. Table I illustrates the summary of the reviewed literature.

A. Prediction of Students' Performance

In 2020, Authors in [29] used Naïve Bayes and J48 techniques for students' academic performance prediction and guided the students by using WEKA tool. They used over 3867 students' records upon of 5 years of Umm Al-Qura University. The results concluded that J48 algorithm achieved the best accuracy of 84.38% while NaiveBayes algorithm gave an accuracy of 46.68%.
Authors in [30] stated that they used educational data mining to predict Semester Grade Point Average (SGPA) of Bachelor of Technology (B. Tech) third-semester computer engineering students. They used a classification based on previous academic performance and student's social conditions. The authors used two classification algorithms were REP Tree and J48 on the data set using 10- fold cross validation to find the relationship between social parameters and students' performance. This was also used for predicting students' performances in the third semester. They applied these algorithms on the data set of 236 at computer engineering students of Punjabi University. Data was collected through a structured questionnaire from students pursuing B. Tech Computer Engineering from the Computer Engineering Department, Punjabi University. They conducted their study on a sample of 260 students having 17 attributes which included social parameters and previous academic performance such as (fathers' and mothers' education, living place during B. Tech, marks obtained in 10th English, marks obtained in 12th English and marks obtained in 10th Math). The study revealed that parent's education affected student's performance. Moreover, second semester performance played an important role for third semester performance. Finally, they found that a J48 Algorithm gave higher accuracy than REP Tress. The accuracy of J48 was 67.37%, while REP Tree was 56.78%.

According to the authors in [31], they worked on predicting the student's final grade based on the information collected in the early stage. Prediction was based on two different training data sets. Each data set contained data of different students in the last four semesters in the period from 2013 to 2015. In their study, the authors used data based on the following parameters (grade, test1, test2, lecture presence, and lab-presence). The research was executed on the basis of two separate experiments. Both experiments had the same goal, as we mention, which was to predict students' final grades. Data mining classification algorithms were applied on both data sets separately, and the goal was to predict students' final marks based on those two data sets. The data was collected from one particular course. The first training data set contained information about a number of students' visits to the lectures and laboratory exercises. The second training data set contained more students' data besides lecture visits which were added as two more parameters. These two parameters were students' results on two tests performed during the semester. The authors concluded that students must be present in one-third of the total number of lectures and laboratory classes to pass a particular exam. The study cited that when the total number of students present in the lecture and laboratory classes was greater than two-thirds of the total number of classes, the students would get a high grade. Also, The authors also used an IBK algorithm, which is the implementation of K-nearest neighbor classifier for the first data set. This is besides J48 classification algorithm, which is an implementation decision tree classifier for another data set. The authors concluded that they found in the Second training that IBK algorithms provided the best performance of 98.58%. Besides, the J48 technique also provided a good performance of 86.40%.

According to [32], the authors used data mining methods for students' academic prediction. They used data based on different parameters, such as teaching material access duration, academic performance for students, including assignments and tests, and discussion forums. In their study, they collected information about students who taken Programming Fundamental and Advanced Operating System courses from August 2014 to May 2015. Then, their study applies three classifiers on the compared, tested, and analyzed data set. The classifiers were Naïve Bayes, Multi-layer perception, and C4.5 (J48). The three classifiers were tested on 38 attributes. They applied ten-fold cross-validation, which means that the data set was randomly divided into ten subsets of the same size. The authors concluded that the Naïve Bayes classifier gives the best overall prediction accuracy than the other two classifiers with 86%.

The study in 2020 [33] by Zainab Mohammed et al. used WEKA tool to predict the academic instructors' performance by Using K-means Clustering and Naive Bayes classifications. They used data set at the UCI website that contained a total of 5820 evaluation scores provided by the university students for the evaluation of the academic instructors' performance and attributes such as instructor's name, course code values, and the course attendance rate. The result found that Naive Bayes classification had an accuracy of 98.86%, while 98.9% for K-means Clustering.

Al Breiki et al. [34] used data mining algorithms to predict the performance of students by using WEKA tool. They analyzed 145 Students data of two academic years (2009- 2010 and 2014-2015) and attributes including student ID, secondary education GPA (/100), and cumulative GPA (/4.00) at the United Arab Emirates University. They applied eight techniques such as Decision Table (DT), propositional rule learning (JRip), Simple Log Regress (SLR), Gaussian Processes Random Tress (GPRT), K-nearest Neighbors (IBK) and Random Forest (RF). The results conclude that Regression gives the best prediction accuracy with 96.98%, followed by Random Forest 96.4%.

In addition, the study [35] by A. Tekin used three prediction techniques of data mining: SVM, ELM, NN and applied to data taken from students who studies computer science and information technology at the end of their 1st, 2nd and 3rd year courses to predict the GPA at graduation. He collected 127 students' records of the Collage of Computer and Instructional Technology enrolled in the Fırat University in Turkey either from 2006 to 2010 or from 2007 to 2011 with the attributes such as grades of students, cultural courses, and student's GPAs. The result of their study highlight that SVM had higher prediction accuracy with a rate of 97.98%, then ELM comes with a 94.8% accuracy rate, and finally NN with the least accuracy rate of 93.76%.

Y. K. Saheed et al. in their study of 2018 [6] applied a data mining technique to predict students' performance based on ID3, J48, and CART algorithms using WEKA tool. They collected 234 student records from the Faculty of Natural Science and Department of Computer Sciences for the years 2013 and 2014 at a private University in Northern part of Nigeria. The result of their study conclude that J48 and CART resulted in the same accuracy of 98.3%, while an ID3 gave 95.9% accuracy in prediction.

Fadhilah Ahmad et al. in their study [36], applied Decision Tree, Naive Bayes, and Rule-Based techniques for predicting the academic performance of first-year bachelor students at the Collage of Informatics and Computing, University Sultan Zainal Abidin, Malaysia. They collected 497 students' records across the span of eight years from 2006/2007 till 2013/2014. The records included data about students such as previous academic records, the background of family and demographics, etc. The results from the research found that the Rule-Based was the best accuracy comparing to the other classification with accuracy rate of 71.3%, while Naive Bayes and Decision Tree found the accuracy of 67.0% and 68.8%, respectively.

In 2018, the study [4] by Alaa Hamoud et al. presented a model based on decision tree algorithms: J48, Random Tree, and REP Tree to predict students' performance. They surveyed students of the Computer Science and Information Technology College in Basrah University. They collected 161 questionnaires and attributes, including academic information, Social Information, Demographic Data, etc. Finally, their result found that the J48 algorithm had the highest accuracy of 62.1% compared to Random Tree and RepTree algorithms of 61.4%, 60.1%, respectively.

In 2020, the study [37] by Abdullah Baz et al. used Naive Byes classifier for predicting students' academic performance at Umm Al-Qura University, based on the final GPA using WEKA tool. The authors collected a dataset consisting of 138 students with 13 attributes. Finally, the results highlight that the Naive Bayes classification had an accuracy of 72.46%.

In 2019, Ramaswami et al. [38] used four data mining techniques that included Logistic Regression, Random Forest, k-Nearest Neighbour and Naïve Bayes. The authors used different techniques in order to improve the prediction accuracy of students' performance by using Python. They collected 240 students of Xorro-Q (Web-based audience interaction tool) from 2016 to 2017 with the attributes such as activity name, activity, test1, test2, and final exam score. The results found that Random Forest had the best accuracy of 74%.

The study [39] analyzed students' data in higher education to predict students' grades and to enhance the students' performance by finding the connection among three main dimensions. The first one was students' activities through e-Learning. The second dimension was teaching manner. And the last dimension is students' results. Their data was collected from the e-Learning system log file and the database of the British University in Egypt. The study's result appeared that the Naive Bayes network had rate of 87.07% prediction accuracy.

In 2017, Uddin, Humam and Nafis, Md Tabrez [40] applied data mining technology for acquiring student performance during their entire semester by using Rapid Miner Studio tool. They used three different clustering techniques: K-Means, K-Medoids, and X-Means for categorizing students. Data set of 94 students of Bachelor of Technology was collected. The batch included 24 attributes such as Aggregate percentage, industry internships, and projects completed. The results of the study showed that X-Means clustering technique gave the best result for students' performance of 86.17% and the accuracy of 81.91% for K-Means, and 84.04% for K-Medoids.

In 2020, Nemomsa et al. in their study [41] used six different classifiers J48, RandomForest, NaiveBayes, BayesNet, JRip, and PART to predict students' academic performance by categorizing student status into dropout/fail, poor, good, excellent, or average performer through predictive modelling by using WEKA tool. They applied the same classifications on two data sets. They collected 6,573 students from AMU student repositories and some data was collected by using questionnaire-based survey. The data collection covered four departments of Computer Science (CS), Water Resource and Irrigation Engineering (WRIE), Water Supply and Environmental Engineering (WSEE) and Hydraulics and Water Resource Engineering (HWRE) of the second-year first semester. For each department, one major course was selected, with several attributes including gender, department, course, course credit hours, grade, semester average GPA, cumulative GPA, and students' status. The result showed that J48 and JRip classifiers produced the highest classification accuracy of J48 is 99.4%, and for JRip, it was 99.3%.

Another study [42] by S. Senthil and W. LIN applied fourteen classification methods to enhance and evolve models that target students' performance prediction and identify the impact attributes. The data set was taken from UCI Machine Learning Repository while contained 33 attributes of 649 students. The study's result highlighted that BayesNet, Multilayer-Perceptron, Simple Logistics Pegasos, KStar, JRip, and Random Forest were the best algorithms for accurate prediction. The algorithm with the highest accuracy was the Simple Logistic of 93.2% followed by Random Forest of 93.1%, and the worst algorithm was IBk with 82.1%.

The study [43] by Alhakami et al., analyzed students' performance and achievement regarding ABET files learning by using Naive Bayes and J48 algorithms. They used a data set consisting of 126 students and some attributes such as attendance, quiz, midterm, and grade at Umm Al-Qura University. The results appeared that the J48 classifier had the best accuracy of 100%.

B. Detecting Students Behavior

The study [44] by Pujianto et al. in 2017 aimed to assist the students at the faculty of literature and the likelihood of their success in adapting to new environments. In Indonesia, it has always been an issue for elected students to join the literature faculty, especially those who don't have linguistic qualifications in high school. Their study applied the Naive Bayesian classifier algorithm to predict those students' level of achievements in Literature Faculty who came from the non-linguistics major in senior high school. The authors used data set based on a survey published online for students from literature faculty all around East Java Province in Indonesia. Some attributes used in this study were the English national exam score and the number of books reads per month. Based on the data analysis by using the NBC algorithm, the authors found that the high school senior students who don't have linguistic traits were also able to join the literature faculty. For the analysis of talent and the national exam score, the results showed that the accuracy of NBC was 70%.

On the other hand, to predict the probability of student's graduation at JIIT University of India. R. Ahuja, and Y. Kankane [14] applied Seven techniques for education data mining. The techniques used include KNN, Naïve Bayes, Random Forest, Logistic Regression and Ctree by using R language. The data set used for this research consisted of 35 attributes of students that included both the academic and non-academic data such as students' grades, student's age, size of family, residential, travelling time from home to school. The data was complied by using college reports. The result of the study concluded that the Ctree and Random Forest algorithm performed much better than other algorithms. The prediction accuracy of the two techniques was 90.37% and 89.47%, respectively.

Another study in 2020 [45] by Hooshyar et al. proposed a novel algorithm called PPP to predict students' performance with learning obstacles by way of procrastination behavior. This was attempted by using eight different classifications such as L-SVM, R-SVM, DT, RF, and NN. The authors collected 242 students and 16 attributes such as open date of an assignment, date of first view of the assignment, and date of assignment submission from the University of Tartu in Estonia. The result showed that PPP algorithm had the maximum accuracy rate of 96%.

C. Enrolment Decision for Students

In 2020, Nurhachita et al., in their study [46], presented a comparison between K-Means and Naïve Bayes clustering methods to use data mining on new students' admission at the Universities Islam in Palembang by using Rapid Miner tool. The authors collected data from 2016 to 2019 of 18930 students with attributes such as students' name, school origin, secondary national examination score, and study programs. The result of their research conclude that the Naive Bayes classifier gave an accuracy of 9.08%.

In 2020, the study [47] by Hanan Mengash aimed to focus on helping universities make acceptance decisions by applying data mining techniques for applicants' academic performance prediction. The study used four classification techniques, including Decision Trees, Support Vector Machines, Artificial Neural Network and Naive Bayes, to predict the students' performance at the end of their school years. The data set contained 2039 students' records of Computer and Information Sciences collage at Princess Nourah Bint Abdulrahman University. The results found that the Artificial Neural Network (ANN) had more than 79% accuracy rate, which makes it better than other classification methods. The Naive Bayes had the worst results.

D. Miscellaneous Studies

In 2020, the study [48] by Fatima Alshareef et al. reviewed the related researches in EDM, including applications and techniques, and identified the best algorithm for each of the EDM applications. The authors had relied on the right prediction accuracy and use it as a guide for identifying effective techniques. Thus, their result conclude that Random Forest and Bayesian were the most algorithms performed effectively for predicting the performance of students. Furthermore, Social Network Analysis gave the best functionality for identifying student behaviors. Both Social Network Analysis and Clustering were the most effective algorithms for student modeling and students' grouping, respectively.

In 2019, Francesco Agrusti et al. [49] tried to collect studies that used EDM methods to predict dropouts students. They selected 73 studies related to this topic to analyze. Their study found six classification techniques that were used in that field. These were: Decision Tree, K-NN, SVM, Bayesian Classification, NN, and Logistic regression. Their study highlight that frequency of the use of Decision tree was 67%, followed by Bayesian Classification at 49 49%, Neural Networks 40%, and Logistic regression 34%.

V. Discussion

A thorough review of the existing research studies shows that there are several algorithms for EDM applications in the context of analyzing students' data to support the educational process. Table II shows a comparison among different research studies based on the highest prediction accuracy depending on the use of the techniques. The results found that the J48 and K-mean are the best effective algorithms in predicting students' performance. The EDM techniques that have achieved the highest usage are Bayesian Classification, followed by Decision tree classifiers, then Logistic regression, Neural Networks, and K-Nearest Neighbour. The study paper

TABLE I. SUMMARY OF RESULTS OF RESEARCH SEEKING STUDENTS PERFORMANCE PREDICTION

Ref.	Objective	Techniques used	Sample size	Best algorithm	Prediction Accuracy	Tool
[29]	Prediction Students Performance	J48 ;NB	38671	J48	84.38%	WEKA
[30]		RT ; J48	260	J48	67.37%	WEKA
[31]		IBK; J48; ZeroR; Part	-	IBK	98.58%	WEKA
[32]		NB; MP ; J48	60	NB	86%	WEKA
[41]		J48; RF ; NB; BN; JRip; PART	6573	J48	99.4%	WEKA
[42]		BN ; MP; SL; SPegasos; KStar; RF;JRip	649	SL	93.2%	WEKA
[39]		NB	3040	NB	87.07%	STATA
[40]		K-Means; K-Medoids ; X-Means	94	X-Means	86.17%	RapidMiner
[35]		NN ; SVM; ELM	127	SVM	97.98%	-
[49]		DT; K-NN; SVM;Bayesian; NN; LR	73	DT	67%	-
[38]		NB; LR; K-NN,; RF	240	RF	74%	Python programming
[34]		SLR; DT; GPRT; IBK; RF;MP;SMOReg;LA	145	LA	96.98%	WEKA
[33]		K-means; NB	5820	K-means	98.9%	WEKA
[36]		DT; NB; Rule-Based	497	Rule-Based	71.3%	WEKA
[6]		ID3; J48; CART	234	J48 and CART	98.3%	WEKA
[4]		J48, RT; REP Tree	161	J48	62.1%	WEKA
[37]		NB	138	NB	72.46%	WEKA
[43]		NB;J48	126	J48	100%	WEKA
[45]	Detecting Students Behavior	PPP; R-SVM; L-SVM; DT; NN; NB; GP; ADB; RF;	242	PPP	96%	-
[14]		NB ; KNN ; RF ; Ctree; LR; Rpart; J48	-	Ctree	90.37%	R programming
[44]		NB	50	NB	70%	WEKA
[46]	Enrollment Decision for Students	NB; K-means	18930	NB	9.08%	RapidMiner
[47]		DT; SVM; NB; ANN	2039	ANN	79%	WEKA

TABLE II. COMPARISON OF THE EDM TECHNIQUES REGARDING ON PREDICTION ACCURACY.

Ref.	Techniques	Highest accuracy appeared
[43]	J48	100%
[46]	NavieBayes(NB)	98.86%
[40]	X-Means	86.17%
[35]	Support Vector Machine (SVM)	97.98%
[14]	Ctree	90.37%
[49]	Decision Tree(DT)	67%
[38]	Random Forest(RF)	96.4%
[34]	Logistic Regression(LA)	96.98%
[45]	Neural Network(NN)	96%
[33]	K-means	98.9%
[36]	Rule-Based	71.3%
[6]	CART	98.3%
[4]	RepTree	61.4%
[6]	Iterative Dichotomiser 3(ID3)	95.9%
[39]	IBK	82.1%
[39]	Simple Logistic	93.27%
[44]	JRip	83.46%
[35]	K-Medoids	84.04%

found that most researchers used two algorithms, Naive Bayes and J48. In addition to its easy implementation and high prediction accuracy, Naive Bayes algorithm deals well with missing data. J48 is another easy to implement algorithm, yet it provides high accuracy results, which explains its frequent use. Besides this, J48 can use both discrete continuous values and has the capability of updating and reasoning. However, it is hard to deal with the absence of data through this technique. However, one should note some of the limitations of these techniques such as their need of large data sets for attaining good accuracy. For further discussion, authors in [50][51] report some advantages and disadvantages of both techniques. In addition, the present paper has identified each tool used in the different research studies and a comparison of those techniques has been tabulated below. Table I shows that 20 out of 23 selected research studies mention the software used; therefore, the results highlight that the most widely used tool was WEKA that attracted 15 out of 20 research, which is approximately 75%. Moreover, using many algorithms in software to identify the prediction accuracy is essential for comparing algorithms and to determine the suitable technique that can be used in the given application.

VI. CONCLUSION AND FUTURE WORK

The EDM would help all the educational stakeholders in several ways. For instance, such tools and techniques could support to improve the students' performance and success in academics, leverage teachers' performance, and support decision-making in institutions. Thus, data mining in higher education would help institutions and educators enhance the educational process effectively. The strength of this review paper lies in using the true prediction accuracy as an indicator to determine the highest effective techniques for each EDM applications of the surveyed studies.

The results of this review paper would be an effective reference for researchers, education providers, educational decision-makers, and others so that they can implement and promote educational data mining more efficaciously. This review paper focuses on further developments in the field of education data mining to support academic advising. Moreover, additional surveys have to be considered for every EDM application by including other standards to specify the best algorithms more accurately.

REFERENCES

[1] J. Jacob, K. Jha, P. Kotak, and S. Puthran, "Educational data mining techniques and their applications," in *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, pp. 1344–1348, IEEE, 2015.

[2] A. Nguyen, L. Gardner, and D. Sheridan, "Data analytics in higher education: An integrated view," *Journal of Information Systems Education*, vol. 31, no. 1, p. 61, 2020.

[3] N. Walia, M. Kumar, N. Nayar, and G. Mehta, "Student's academic performance prediction in academic using data mining techniques," *Available at SSRN 3565874*, 2020.

[4] A. Hamoud, A. S. Hashim, and W. A. Awadh, “Predicting student performance in higher education institutions using decision tree analysis,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, pp. 26–31, 2018.

[5] B. Al Breiki, N. Zaki, and E. A. Mohamed, “Using educational data mining techniques to predict student performance,” in *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, pp. 1–5, IEEE, 2019.

[6] Y. Saheed, T. Oladele, A. Akanni, and W. Ibrahim, “Student performance prediction based on data mining classification techniques,” *Nigerian Journal of Technology*, vol. 37, no. 4, pp. 1087–1091, 2018.

[7] B. K. Baradwaj and S. Pal, “Mining educational data to analyze students’ performance,” *arXiv preprint arXiv:1201.3417*, 2012.

[8] R. Sumitha, E. Vinothkumar, and P. Scholar, “Prediction of students outcome using data mining techniques,” *International Journal of Scientific Engineering and Applied Science (IJSEAS)–Volume-2, Issue-6*, 2016.

[9] E. Alyahyan and D. Düştegör, “Predicting academic success in higher education: literature review and best practices.,” *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, pp. 1 – 21, 2020.

[10] B. Bakhshinategh, O. R. Zaiane, S. ElAtia, and D. Ipperciel, “Educational data mining applications and tasks: A survey of the last 10 years,” *Education and Information Technologies*, vol. 23, no. 1, pp. 537–553, 2018.

[11] A. B. Zoric, “Benefits of educational data mining,” Sep 2019. Copyright - Copyright Varazdin Development and Entrepreneurship Agency (VADEA) Sep 19/Sep 20, 2019; Last updated - 2020-01-22.

[12] H. Almarabeh, “Analysis of students’ performance by using different data mining classifiers,” *International Journal of Modern Education and Computer Science*, vol. 9, no. 8, p. 9, 2017.

[13] A. B. Zoric, “Benefits of educational data mining,” Sep 2019. Copyright - Copyright Varazdin Development and Entrepreneurship Agency (VADEA) Sep 19/Sep 20, 2019; Last updated - 2020-01-22.

[14] R. Ahuja and Y. Kankane, “Predicting the probability of student’s degree completion by using different data mining techniques,” in *2017 Fourth International Conference on Image Information Processing (ICIIP)*, pp. 1–4, IEEE, 2017.

[15] E. N. Azizah, U. Pujianto, E. Nugraha, *et al.*, “Comparative performance between c4. 5 and naive bayes classifiers in predicting student academic performance in a virtual learning environment,” in *2018 4th International Conference on Education and Technology (ICET)*, pp. 18–22, IEEE, 2018.

[16] S. Roy and A. Garg, “Analyzing performance of students by using data mining techniques a literature survey,” in *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, pp. 130–133, IEEE, 2017.

[17] R. B. Sachin and M. S. Vijay, “A survey and future vision of data mining in educational field,” in *2012 Second International Conference on Advanced Computing & Communication Technologies*, pp. 96–100, IEEE, 2012.

[18] R. Baker *et al.*, “Data mining for education,” *International encyclopedia of education*, vol. 7, no. 3, pp. 112–118, 2010.

[19] S. M. Thakrar, N. Jadeja, and N. Vadher, “Educational data mining: A review.,” *IUP Journal of Information Technology*, vol. 14, no. 1, 2018.

[20] C. Anuradha, T. Velmurugan, and R. Anandavally, “Clustering algorithms in educational data mining: a review,” *International Journal of Power Control and Computation*, vol. 7, no. 1, pp. 47–52, 2015.

[21] S. Srivastava, “Weka: a tool for data preprocessing, classification, ensemble, clustering and association rule mining,” *International Journal of Computer Applications*, vol. 88, no. 10, 2014.

[22] M. Chunqiao, “Student performance early warning based on data mining.,” *International Journal of Performability Engineering*, vol. 15, no. 3, pp. 822 – 833, 2019.

[23] K. Saravanapriya, “A study on free open source data mining tools,” *International Journal of Engineering and Computer Science*, vol. 3, no. 12, pp. 9450–9452, 2014.

[24] F. Ali, D. Bhatt, T. Choudhury, and A. Thakral, “A brief analysis of data mining techniques,” in *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, pp. 752–758, IEEE, 2019.

[25] M. Figueiredo, L. Esteves, J. Neves, and H. Vicente, “A data mining approach to study the impact of the methodology followed in chemistry lab classes on the weight attributed by the students to the lab work on learning and motivation,” *Chem. Educ. Res. Pract.*, vol. 17, pp. 156–171, 2016.

[26] P. Shruthi and B. Chaitra, “Student performance prediction in education sector using data mining,” 2016.

[27] F. A. Ibrahim and O. A. Shiba, “Data mining: Weka software (an overview),” *Journal of Pure and Applied Sciences*, vol. 18, no. 3, 2019.

[28] S. Križanić, “Educational data mining using cluster analysis and decision tree technique: A case study,” *International Journal of Engineering Business Management*, vol. 12, p. 1847979020908675, 2020.

[29] H. Alhakami, T. Alsubait, and A. Aljarallah, “Data mining for student advising,” *International Journal of Advanced Computer Science and Applications Science (IJSEAS)–Volume-11, Issue-3*, 2020.

[30] W. Singh and P. Kaur, “Comparative analysis of classification techniques for predicting computer engineering students’ academic performance,” *International Journal of Advanced Research in Computer Science*, vol. 7, no. 6, 2016.

[31] M. Ilic, P. Spalevic, M. Veinovic, and W. S. Alatresh, “Students’ success prediction using weka tool,” *Infoteh-Jahorina*, vol. 15, pp. 684–688, 2016.

[32] A. Mueen, B. Zafar, and U. Manzoor, “Modeling and predicting students’ academic performance using data mining techniques,” *International Journal of Modern Education and Computer Science*, vol. 8, no. 11, p. 36, 2016.

[33] Z. M. Ali, N. H. Hassoon, W. S. Ahmed, and H. N. Abed, “The application of data mining for predicting academic performance using k-means clustering and naïve bayes classification.,” *International Journal of Psychosocial Rehabilitation*, vol. 24, no. 3, pp. 2143 – 2151, 2020.

[34] B. Al Breiki, N. Zaki, and E. A. Mohamed, “Using educational data mining techniques to predict student performance,” in *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, pp. 1–5, IEEE, 2019.

[35] A. Tekin, “Early prediction of students’ grade point averages at graduation: A data mining approach,” *Eurasian Journal of Educational Research*, vol. 54, pp. 207–226, 2014.

[36] F. Ahmad, N. H. Ismail, and A. A. Aziz, “The prediction of students’ academic performance using classification data mining techniques,” *Applied Mathematical Sciences*, vol. 9, no. 129, pp. 6415–6426, 2015.

[37] A. Baz, F. Alshareef, E. Alshareef, H. Alhakami, and T. Alsubait, “Predicting students’ academic performance using naïve bayes,” *International Journal of Computer Science and Network Security*, vol. 20, no. 4, 2020.

[38] G. Ramaswami, T. Susnjak, A. Mathrani, J. Lim, and P. Garcia, “Using educational data mining techniques to increase the prediction accuracy of student academic performance,” *Information and Learning Sciences*, 2019.

[39] E. Abou Gamie, S. Abou El-Seoud, M. Salama, and W. Hussein, “Multi-dimensional analysis to predict students’ grades in higher education,” *International Journal of Emerging Technologies in Learning (iJET)*, vol. 14, no. 02, pp. 4–15, 2019.

[40] H. Uddin and M. T. Nafis, “Students academic performance using partitioning clustering algorithms,” *International Journal of Advanced Research in Computer Science*, vol. 8, no. 5, 2017.

[41] G. Nemomsa, D. P. Sharma, and A. Mulugeta, “Predictive modeling for student performance analytics through data mining techniques,” *IUP Journal of Computer Sciences*, vol. 14, no. 1, 2020.

[42] S. Senthil and W. M. Lin, “Applying classification techniques to predict students’ academic results,” in *2017 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC)*, pp. 1–6, IEEE, 2017.

[43] H. H. Alhakami, B. A. Al-Masabi, and T. M. Alsubait, “Data analytics of student learning outcomes using abet course files,” in *Science and Information Conference*, pp. 309–325, Springer, 2020.

[44] U. Pujianto, E. N. Azizah, and A. S. Damayanti, “Naive bayes using to predict students’ academic performance at faculty of literature,” in *2017 5th International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, pp. 163–169, IEEE, 2017.

[45] D. Hooshyar, M. Pedaste, and Y. Yang, “Mining educational data to predict students’ performance through procrastination behavior,” *Entropy*, vol. 22, no. 1, p. 12, 2020.

[46] N. Nurhachita and E. S. Negara, “A comparison between naïve bayes and the k-means clustering algorithm for the application of data mining on the admission of new students,” *Jurnal Intelektualita: Keislaman, Sosial dan Sains*, vol. 9, no. 1, pp. 51–62, 2020.

[47] H. A. Mengash, “Using data mining techniques to predict student performance to support decision making in university admission systems,” *IEEE Access*, vol. 8, pp. 55462–55470, 2020.

[48] F. Alshareef, H. Alhakami, T. Alsubait, and A. Baz, “Educational data mining applications and techniques,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, 2020.

[49] F. Agrusti, G. Bonavolontà, and M. Mezzini, “University dropout prediction through educational data mining techniques: A systematic review,” *Journal of e-Learning and Knowledge Society*, vol. 15, no. 3, pp. 161–182, 2019.

[50] S. Roy and A. Garg, “Analyzing performance of students by using data mining techniques a literature survey,” in *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, pp. 130–133, 2017.

[51] J. Charoenpong, B. Pimpunchat, S. Amornsamankul, W. Triampo, and N. Nuttavut, “A comparison of machine learning algorithms and their applications.,” *International Journal of Simulation–Systems, Science & Technology*, vol. 20, no. 4, 2019.